The MSCL Analyst's Toolbox: Statistical analysis of microarray data

Jennifer J. Barb, Zoila G. Rangel and Peter J. Munson

Mathematical and Statistical Computing Laboratory, Center for Information Technology, National Institutes of Health.
12 South Dr. Bldg 12A, Room 2001C, Bethesda, MD 20892
February 2010

Contact information:

Jennifer J. Barb, Ph.D.
Bldg 12A Room 2001C
301-435-9232
barbj@mail.nih.gov

Zoila G. Rangel, M.S.
Bldg 12A Room 2001
301-402-9712
rangelzg@mail.nih.gov

Peter J. Munson, Ph.D.
Bldg 12A Room 2039
301-496-2972
munson@helix.nih.gov

**Last updated February 2010

Table of Contents

Chapter 1:
Obtaining and installing the Toolbox

Description of User's Guide
This book is provided to assist in the use of the *MSCL Analyst's Toolbox* software for statistical analysis of microarray data developed in the JMP statistical discovery software. This book provides an overview of running multiple scripts of the toolbox however it does not provide an in depth explanation of each script. If there are other questions or problems please contact either Jennifer Barb or Peter Munson for further details.

How to obtain a copy of JMP
If you are an NIH employee you may obtain a copy of JMP by contacting your IT desktop support for your department. Look up your contact at
http://sdp.cit.nih.gov/information/contact_lookup.asp
If you are not an NIH employee, you may go to JMP's main website to look into obtaining a copy of the JMP software:
www.jmp.com

How to obtain a copy of the MSCL Analyst's toolbox
You may download the toolbox from the MSCL's website here:
http://affylims.cit.nih.gov/MSCLtoolbox/ (inside NIH) or here
http://abs.cit.nih.gov/MSCLtoolbox/ (outside NIH)
The scripts are available for the Mac, PC, and Linux. They are provided as a .zip file and can be downloaded and saved anywhere on your local drive. If you are running Windows, please see the directions below on how to install the customized MSCLToolbox toolbar functionality.

How to install the toolbox Menu bar (Windows only) and how to run the scripts (Mac and Linux)

Running the scripts in JMP 5 (Windows) and JMP 5-8 (Mac and Linux)

1. Download the MSCLToolbox.zip file and extract the files in the zipped folder.
2. Place the file anywhere on your hard drive so that you know where it is.

** this file contains all annotation files for many chips, we suggest deleting all annotation files that you will not be using as these are large files and they will fill up your hard drive. You will want to do this step before extracting all files from the zip folder. So select the annotation files that you don't need inside the zip dialog and then choose to delete selected items. These files have the following suffix in their name "_annot.JMP"

3. Launch JMP session
4. In order to run a script, you will need to open a data table.
5. Open the particular script that you are interested in running and choose "Run Script" under the Edit menu.
6. You will need to have both the script and data table open in order to run it a script from the toolbox.

Setting the Toolbox up as a menu bar item in JMP 7 and JMP8 for Windows (if you are running JMP 8, you should be sure to have the JMP 8.02 updater installed as there are many problems with JMP8 without the updater)

A. IF YOU ARE INSIDE THE NIH NETWORK AND CAN SEE THE FOLLOWING LINK, you will install the "MSCLToolboxShare Menu Toolbar":

\\mscltoolbox.cit.nih.gov\mscltoolbox

In order to test if you can see this file share location in Windows Explorer:

1. Open "My Computer" from the Start Menu
2. Double Click "C:\" drive
3. In the address bar, type the above location
4. If you are able to open this location, then YOU DO HAVE ACCESS TO THIS LOCATION AND YOU WILL PROCEED WITH INSTALLING THE "Share Menu Toolbar"
5. Skip B below and proceed to the next section stating "Launch JMP7 or JMP8"

B. IF YOU ARE OUTSIDE OF NIH OR CAN NOT VIEW THE LOCATION ABOVE, you will install the "MSCLToolboxCdrive Menu Toolbar"

1. Download the MSCLToolbox from the http://abs.cit.nih.gov/MSCLtoolbox/ link. Extract the MSCLToolbox.zip file. Usually this happens automatically if you click on the .zip file. If it doesn't, then right click the zip file and choose to extract files from it.

2. Drag the folder named "MSCLToolbox" into the C:\ directory and continue reading the next section.

The path to the toolbox should now be:

C:\MSCLToolbox

** this file contains all annotation files for many chips, we suggest deleting all annotation files that you will not be using because these are large files and they will fill up your hard drive. You will want to do this step before extracting all files from the zip folder. So select the annotation files that you don't inside the zip dialog and then choose to delete selected items. These files have the following suffix in their name "_annot.JMP"

1. Launch JMP 7 or JMP 8
2. Go to Edit menu
3. Click Customize
4. Choose Menus and Toolbars
5. Right Click the "Main Menu" tab in the window on the left. Choose Import Menu Archive.

6. Decide which menu bar type you are based on the directions above from parts A or B:
   a. FOR **"MSCLToolboxShare Menu Toolbar**" browse and open the <MSCLToolbox_mscltoolboxShare.jmpmenu> file
   b. FOR "**MSCLToolboxCdrive Menu Toolbar**" browse and open the <MSCLToolboxJMP7_cDrive.jmpmenu> file

7. Click Apply in the upper left hand corner window and then click the "X" to close it.
8. The toolbox should now appear in your menu bar and should run all scripts.
9. Test any script with a table to make sure that it works.

REMOVING THE TOOLBOX FROM THE TOOLBAR

1. Open JMP session.
2. Go to Edit tab
3. Click "Customize"
4. Choose "Revert to Factory Defaults"
5. This will set the menu toolbar back to factory settings.

If you are running the "MSCLToolboxCdrive Menu Toolbar" type of toolbar and you need to update a script in the toolbar

1. Download the updated script.
2. Drag and drop the script to where the toolbox menu is. Please be sure to keep the script in the desired path where it already exists. Example:

If you are updating the 3-TransformData.jsl script, then drag and drop the script to the directory just under the MSCLToolbox folder.
If you are updating a statistical script, ANOVA1.jsl, then drag and drop the script to the directory just under the MSCLToolbox/Statistics folder
note, if you are running the **“MSCLToolboxShare Menu Toolbar” type of toolbar, scripts are updated automatically on the share location and you do not need to update

If you experience problems, please email:
Jennifer Barb at barbj@mail.nih.gov or
Peter Munson at munson@helix.nih.gov

Some important side notes about the Toolbox:

1. Under the View tab in the toolbar, it is always good practice to keep the Log view checked. This will keep the log open at the bottom of the user window. If a script crashes for some reason, then the script will be printed in a jumbled form in the log window. You can also copy what is in the Log window and send it to Jennifer Barb. This can give an idea as to why the script crashed.

2. The script acts upon only the top-most table, if multiple tables are open. You can view the table list under the Windows tab. Be certain that the current data is checked.

How to obtain a copy of Affymetrix Expression Console software

Expression Console™ software supports Probe Set summarization and CHP file generation for both 3' Expression (e.g. GeneChip® Human Genome U133 Plus 2.0 Array) and Exon Arrays (e.g. GeneChip Human Exon 1.0 ST Array). The Expression Console workflow provides the user with a choice of the more commonly used Probe Set summarization algorithms. The algorithms offered include:

- MAS5 Statistical algorithm
- Probe Logarithmic Intensity Error Estimation (PLIER)
- Robust Multichip Analysis (RMA)

Additionally, GeneChip data QC is a key component to this workflow and is augmented by a variety of visualization and graphing tools provided by the Expression Console™ software.

Download information and instructions here:
http://www.affymetrix.com/products/software/specific/expression_console_software.affx

Chapter 2:

Introduction to JMP

Introduction to JMP: The Statistical Discovery Software

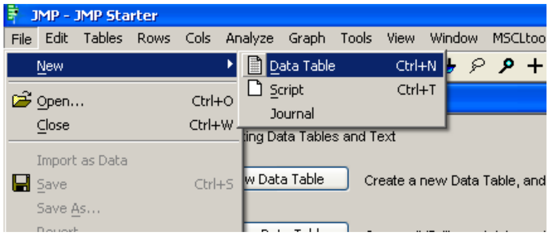
In this chapter the basics of how to use, interact, manipulate and navigate through the JMP statistical discovery software will be briefly introduced. After reading through this chapter, the user should be able to build data tables, import and export data, navigate around their data, manipulate columns and column formulas, execute common table functions, create, save and export graphics and select and link graphics with data tables.

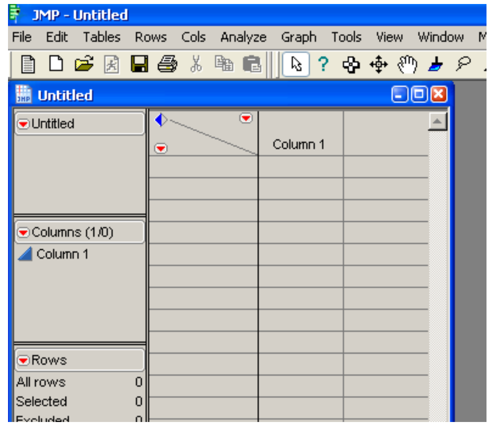
The user should have a solid feel of the power of the JMP software and should have an understanding of the most commonly used capabilities within JMP. Many of the menu items within JMP will be covered and the user should be sure to feel comfortable with navigating within JMP before moving on to more advanced analysis chapters.

The following topics to be covered are outlined below:

A. Creating a datatable
B. Data Table Format
    1. Column options: Label, Hide, Formula
    2. Rows options: Label, Hide, Color
    3. Table Properties
C. JMP tables vs. Excel spreasheets
D. How to Import and Export Data
E. Viewing/Selecting/Analyzing data: Histogram, Scatter Plot, Fit a line, Selecting Tools
F. Table functions: Summary, Subset, Join by matching columns
G. Saving and exporting graphs

A. **Create a Data table**
To create a datatable select File→New→Data Table



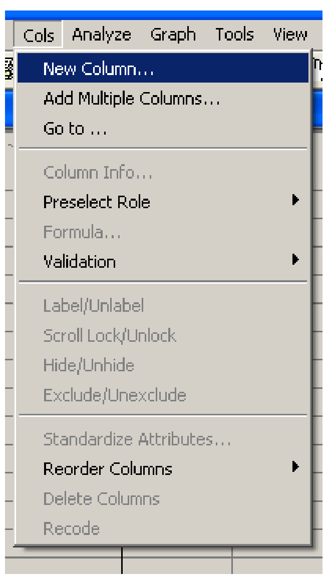


The resulting table will have no rows and one column. The next step is to add columns to be populated with data.

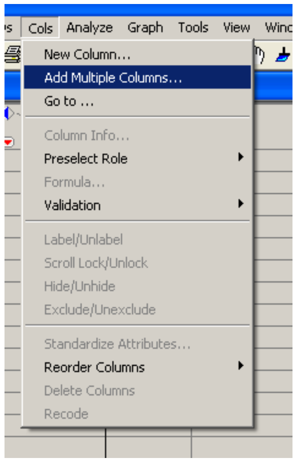
1. Add Columns

   There are several ways to add columns to a data table but we will cover three

   - To add one column select **Cols→New Columns**



   - The second way to add a column is by selecting **Cols→Add Multiple Columns.** Using *Add Multiple Columns* allows the user to chose the number of columns to be add.

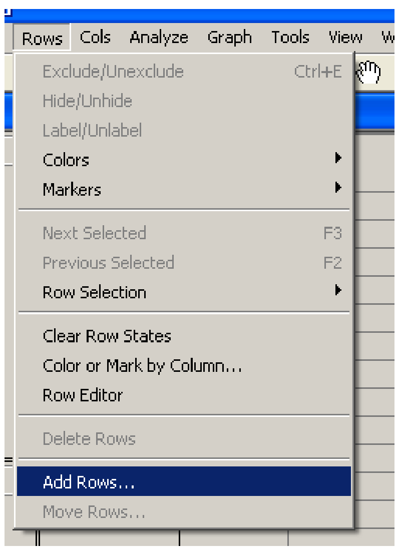


- A third way to add a column is to click in the space after the last existing column header.

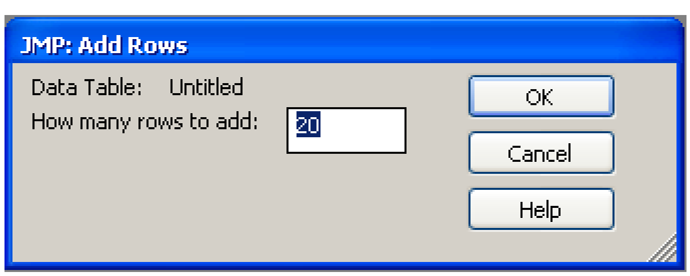
2. Add Rows

   Two ways to add rows to a datatable

   - To add a row select Rows → Add Rows



Then enter the number of rows needed.

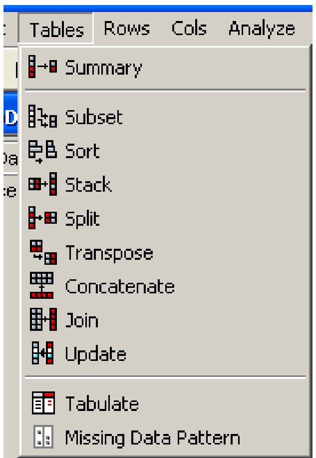


- A row can also be added by double clicking after the last existing row.

3. Menus

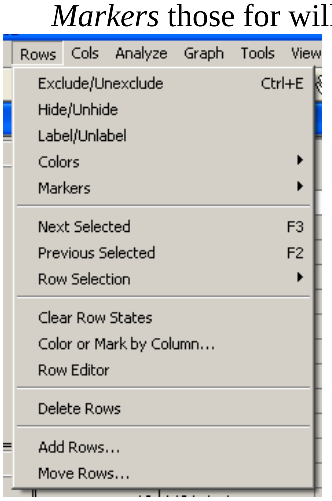
- Tables

Several commonly used options under *Tables* are *Sort*, *Summary*, *Subset* and *Join*. *Summary* summarizes data by creating a table which reports the summarized information on the requested columns. *Join* may be one of the features that set JMP apart from other software. *Join* takes two tables who share a common column and combines them.



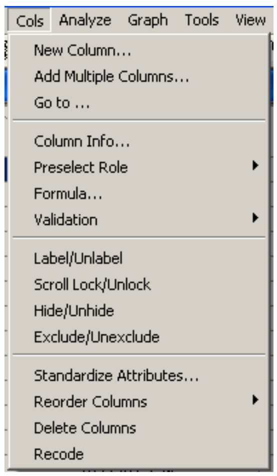
- Rows

Since the most commonly used options are *Hide*, *Label*, *Colors*, *Markers* those for will be explained later in this chapter.



- Cols Menu

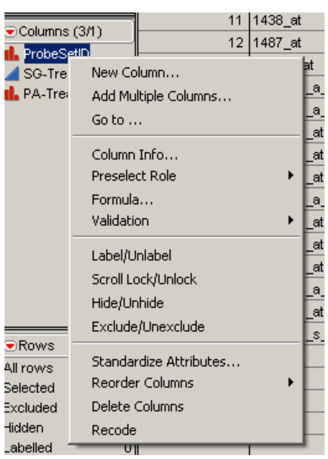
Since the most commonly used options are *Hide*, *Label*, *Formula*, *Column Info*, those four options will be explained later in this chapter.



Data Table Format

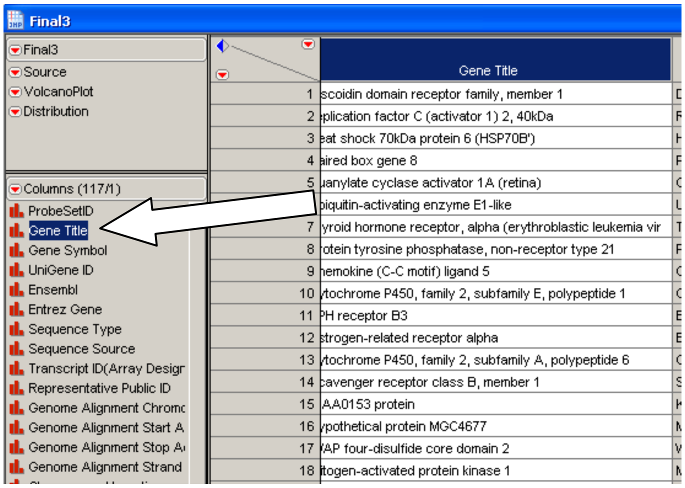
When working in JMP, data is in a data table. So the following section focuses on common operations that will aid in getting the most out of your data.

1. Column options
   Shown below are all the column options available.



- Label

By default row numbers are used as labels when you hover over a point on any visualization tool. To label points with a particular labeling column, select a column by clicking on the column name listed on the left hand side of the data table as shown below



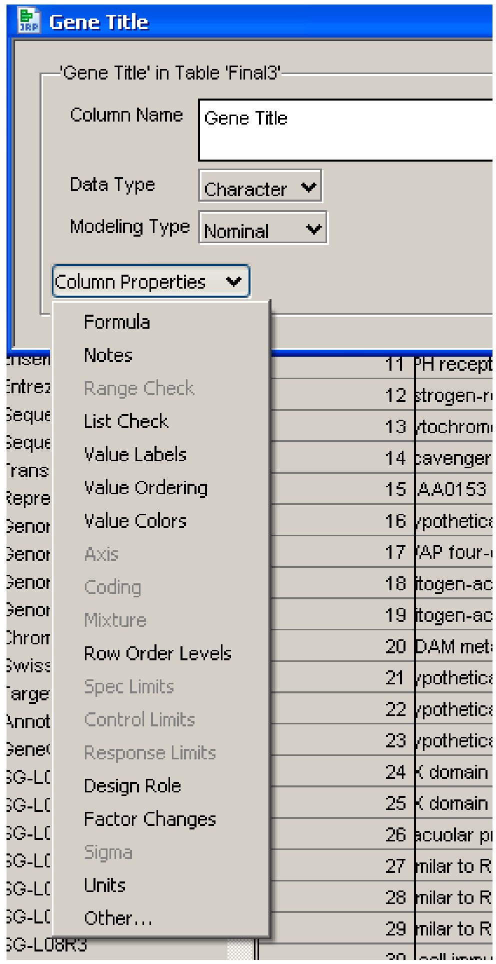
Then select **Cols→Label.**

- Hide

Hide is helpful when the number of columns is unmanageable. To hide columns click on column name to be hidden right click and select hide.

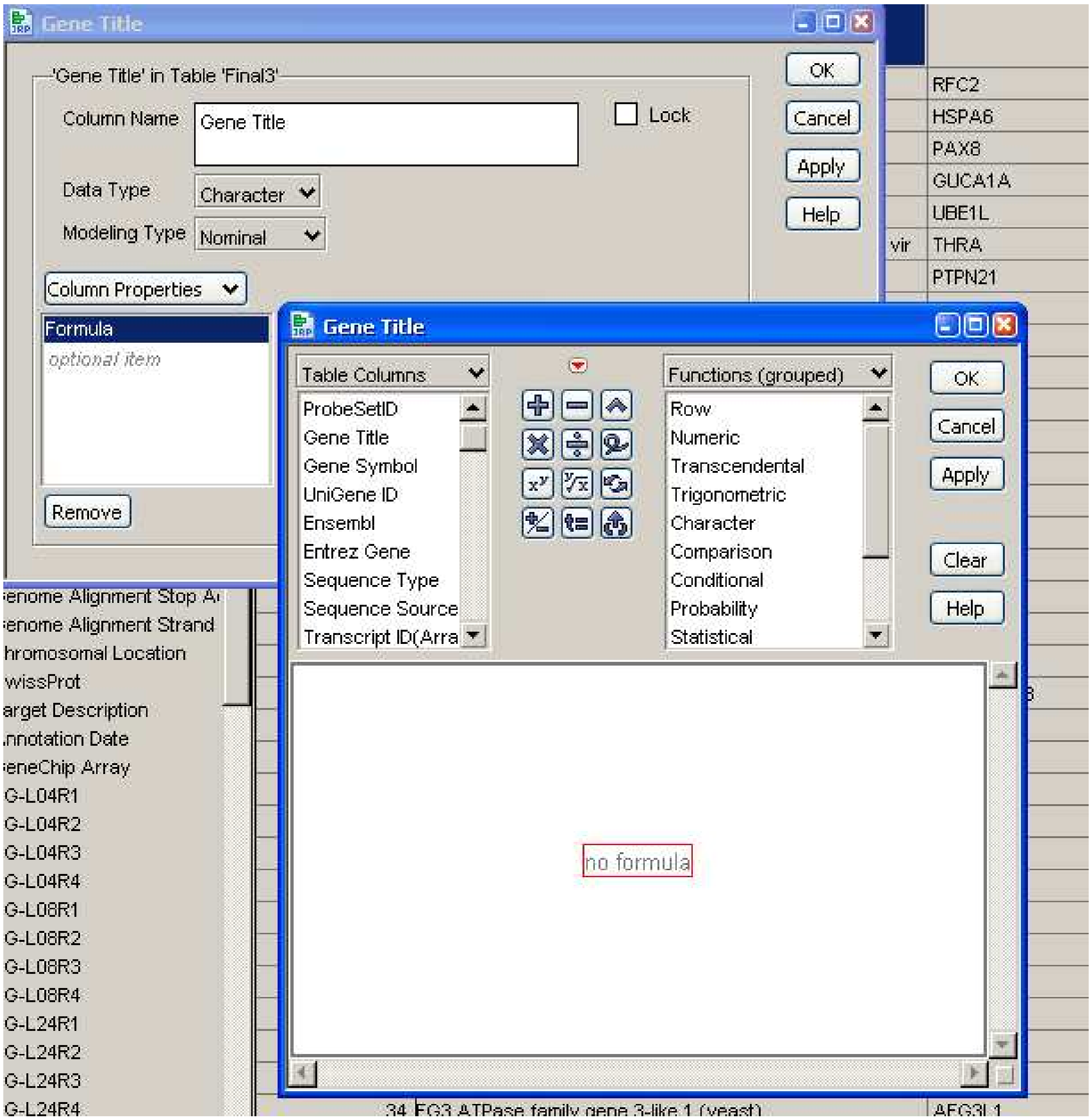
- Column info

Column info allows you to set the data type, modeling type and set column properties.



- Formula

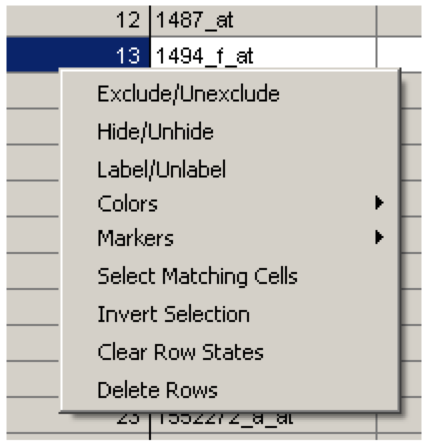
To set a column formula select
Cols → Column Info →Column Properties → Formula →Edit Formula
OR
Hover over column header and right click



There are several operators that can be used to make formulas. The most common operators are in the middle of the dialog box and the rest listed under *Functions*.

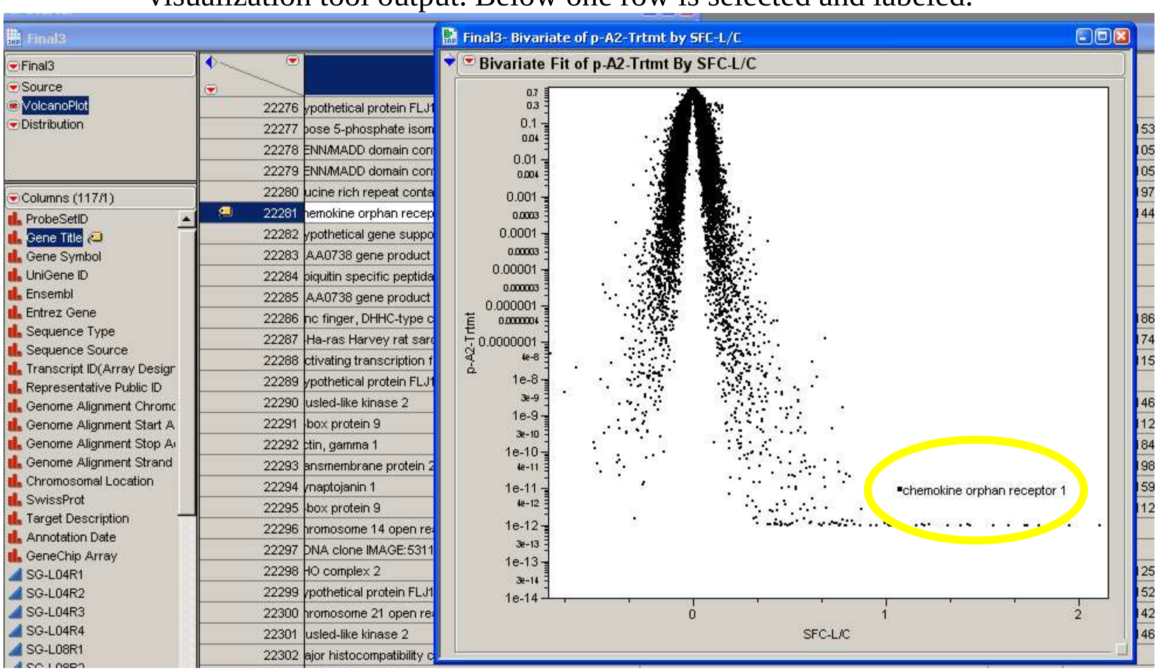
2. Row properties

Shown below are all the column options available.



- Label

Label under the Rows tab allows you to label selected rows or all rows on any visualization tool output. Below one row is selected and labeled.

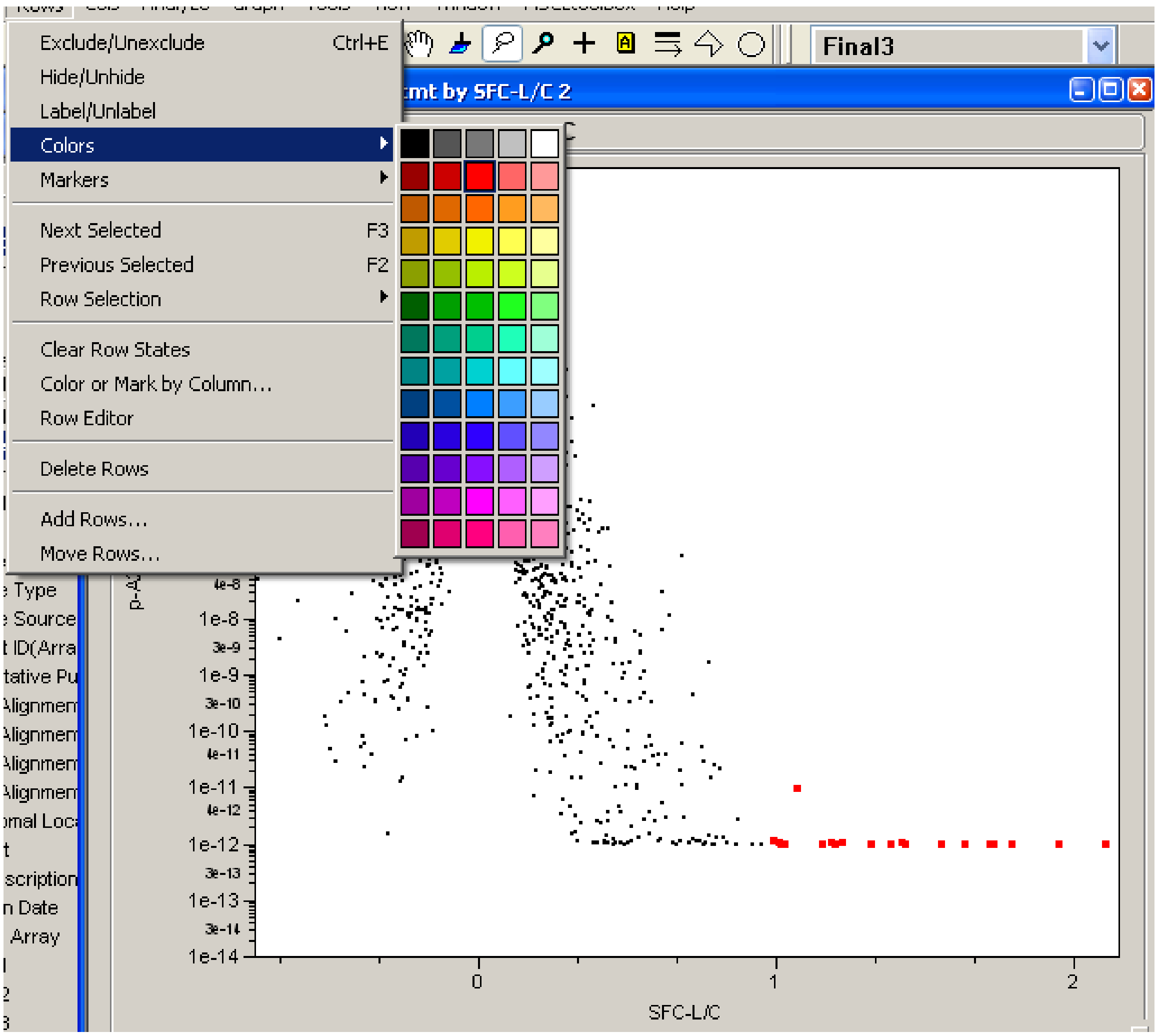


- Hide

Hide is helpful to focus on the observations of interest and hide all others. To hide observations select the rows to be hidden by clicking on the row number then select **Rows→Hide**

- Color and Markers

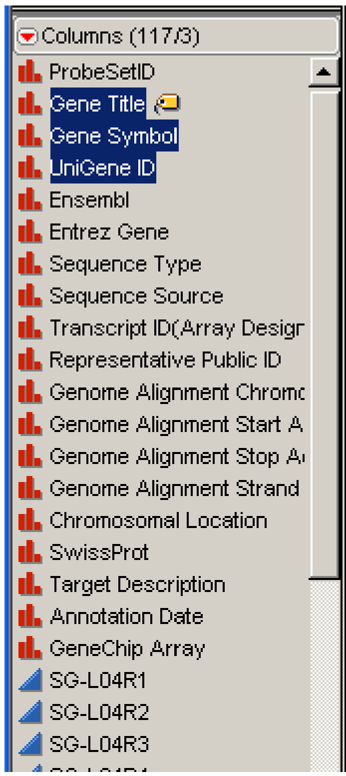
First select the row(s) of interest then select **Rows→Colors** then select a color.

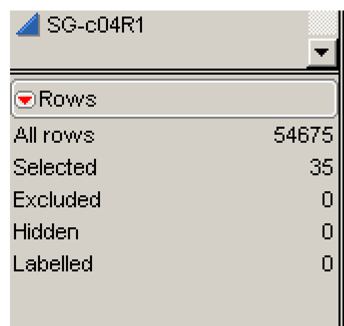


3. Table properties

Below is a column panel (on the left) which summarizes the column information. The numbers in parenthesis are the number of columns/ the number of selected columns. In this example there are 117 columns of which 3 are selected.

The Row panel below (on the right) summarizes row information. In this example there 54,675 rows and 35 of those rows are selected.





A. JMP tables vs. Excel spreadsheets

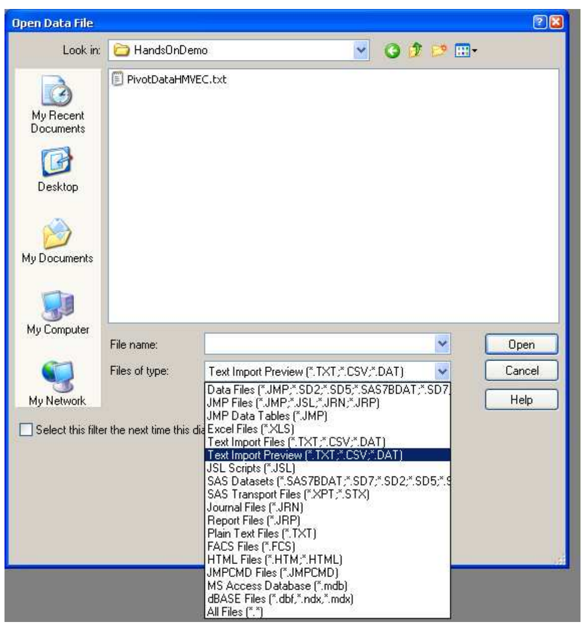
Since Excel is commonly used software there are several ways that data in JMP differs; here are some facts about the way data is handled in JMP.

1. Fixed use of rows and columns
2. Rows are “cases”, columns are “variables”
3. Rows are “ records”, columns are “fields”
4. Rows are “data points”, columns are “dimensions”
5. Cells contain only data (no formulas)
6. Columns may contain formulas
7. Analysis platforms use data from the current Table

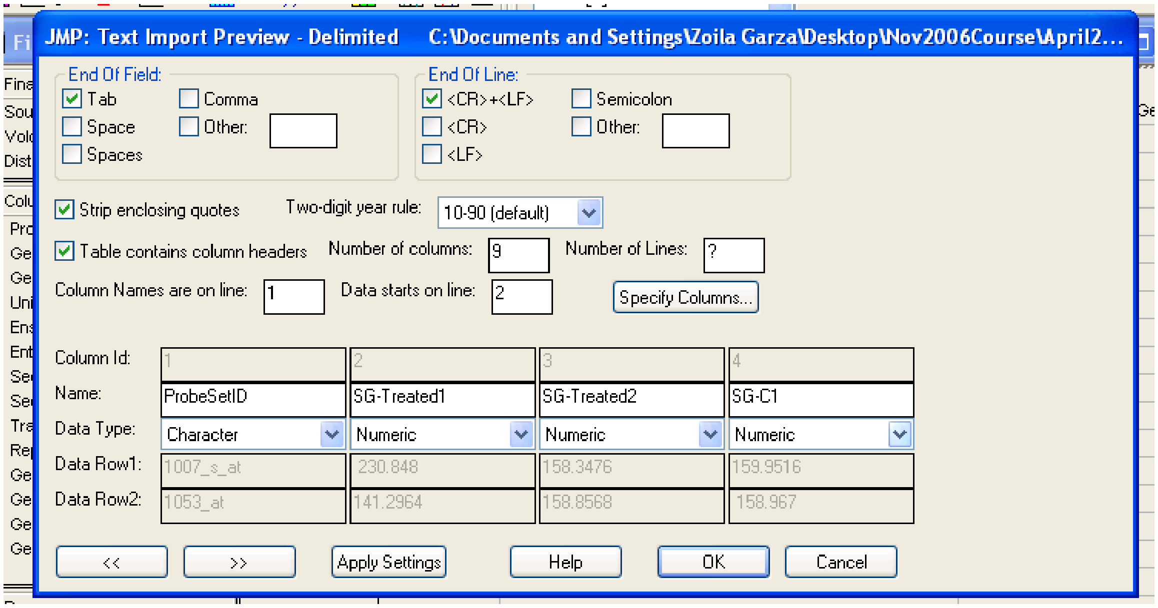
B. How to Import and Export Data

1. Import data using *Text Import Preview*

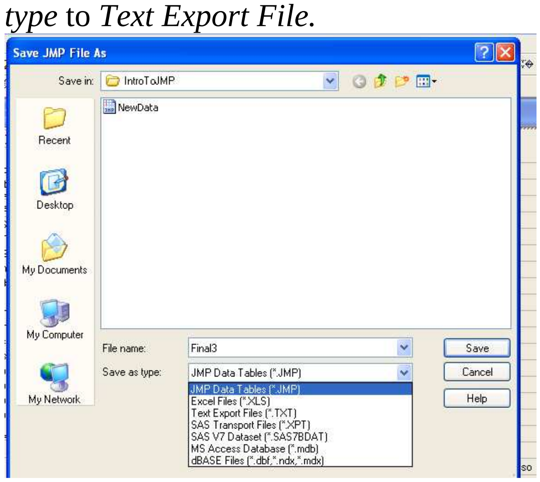
   To import select **File → Open** then change files of type to *Text Import Preview* as shown below.



The following dialog box will appear. There are several options to verify. The section labeled End of Field is important. The section at the bottom allows you to preview the data. The first and second observations/rows are displayed. Verify that the data appears as it should and click OK



2. Exporting data
   To export your data from JMP there are several options. Here are a few in detail:
   a. Save as .txt
      To save as a .txt select **File → Save As** and then chance the *Save as type* to *Text Export File.*



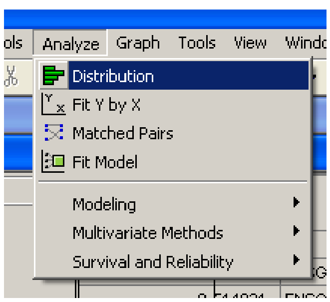
   b. Save as .xls
      Similar to .txt, in order to save as a .xls simply change the *Save as type* to *Excel File*
   c. Copy and paste

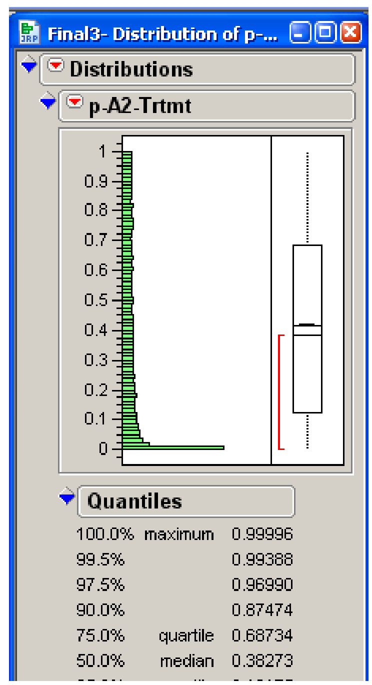
The last way of exporting data is by copying the whole data table and pasting it into any other package. This method is not recommended since there are many ways to err unknowingly.

C. Viewing/Selecting/Analyzing data

1. Histogram – Analyze/Distribution
   Histogram allows you to see a summarized view of data. The result will include a box plot along the side of the distribution, quantiles and summary statistics.

   Select **Analyze → Distribution**

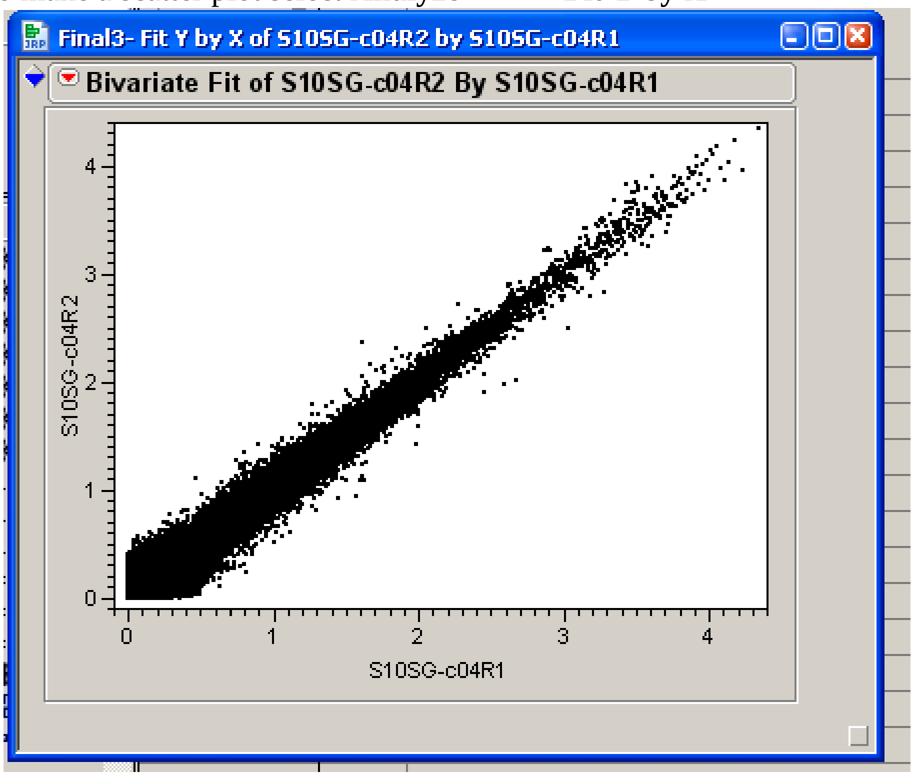




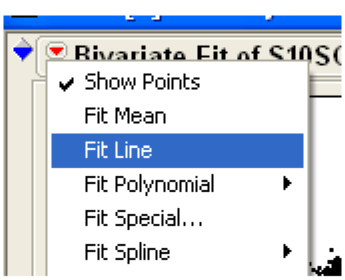
2. Scatter plot – Analyze/Fit Y by X

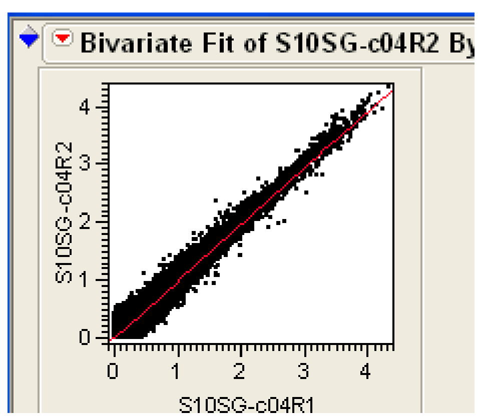
Fit Y by X provides a graph commonly called a scatter plot. There are several options available within a scatter plot to highlight points to aid visualization. The two discussed in this section are *Fit Line* and *Fit Special*.

To make a scatter plot select **Analyze --> Fit Y by X**

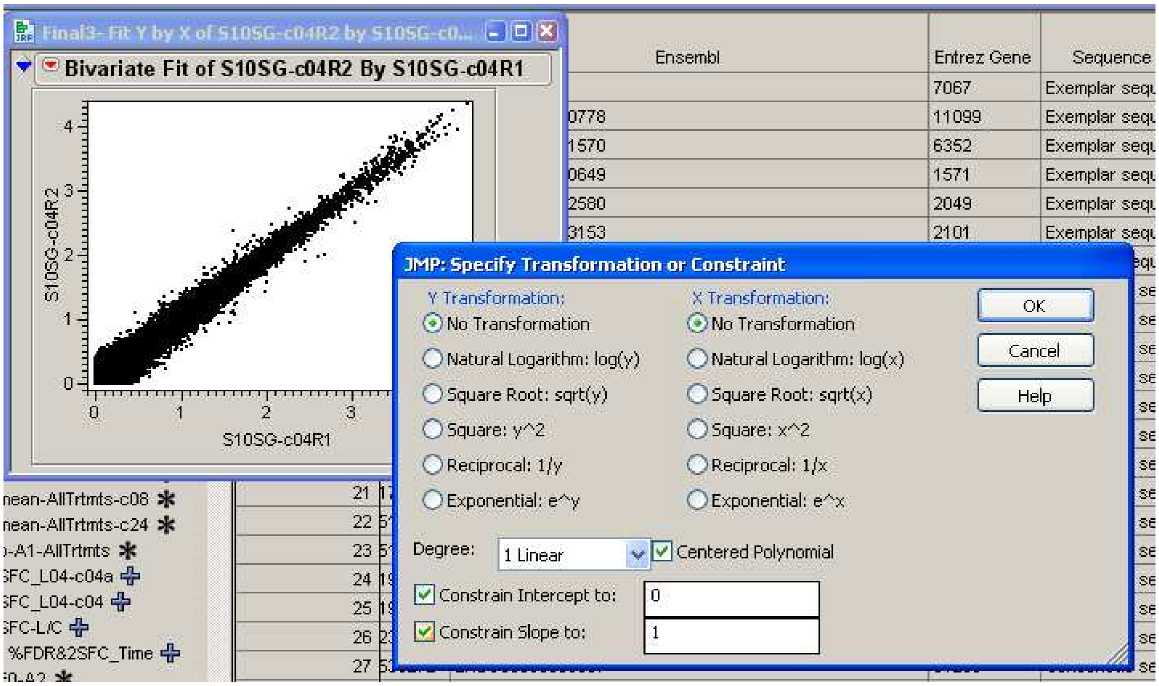


- Fit a line
  To fit a line to the data with the plot click the red triangle and select *Fit Line* as shown below on the left. The results will appear as in the graph on the left.

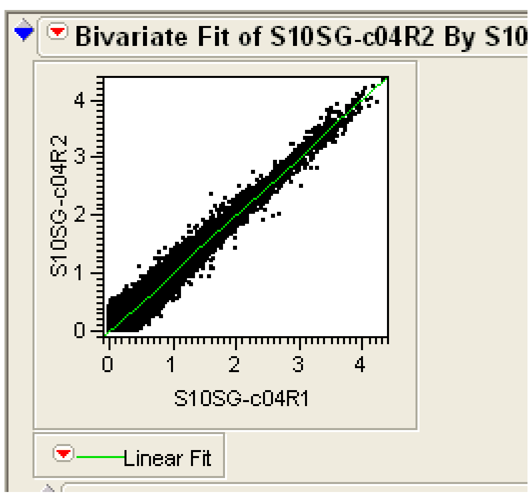




- Add the line of identity
  To add the line of identity click the red triangle and choose *Fit Special* then check the two boxes (Constrain Intercept to: , Constrain Slope to:) as shown below.

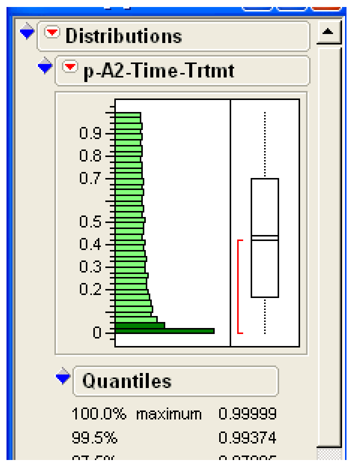


The resulting graph has the line of identity.



3. Tools – arrow, lasso, hand, paint brush
Tools such as the arrow, lasso, hand and paint brush can be helpful to explore data; especially for selecting interesting observations/points. These tools can be found by selecting **Tools**

4. Select from a histogram
There are several ways to select rows or observations within a histogram. The arrow will allow you to click on each bar in graph. Use shift click to select more than one bar. In the histograms two bars have been selected using the arrow.

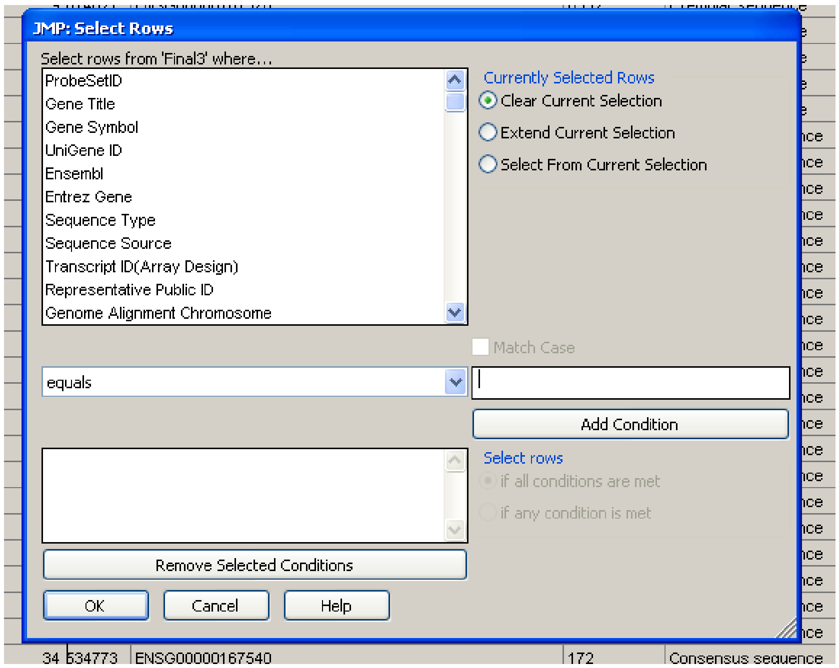


5. Select from Y by X plot

One option for selecting points of interest on an Y by X plot is using the lasso. Select **Tools → Lasso** and encircle the points of interest. The points will then be selected on both the graph and in the data table.

6. Rows /row selection/select where

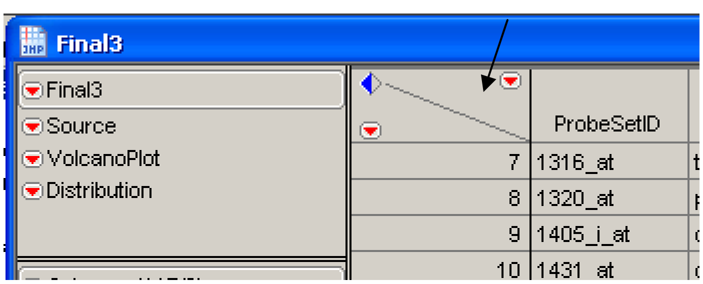
When in search of one particular observation selecting **Rows→Row Selection → Select Where** will allow you to search the entire datatable. The dialog box below prompts for selection criteria



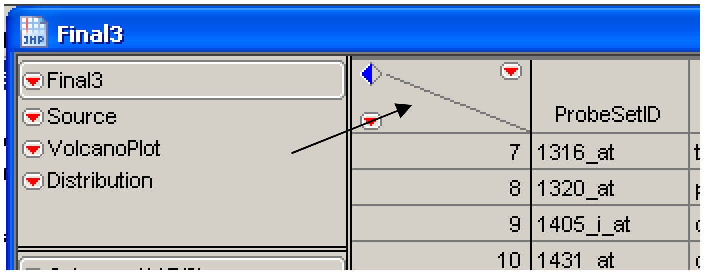
7. Clear a row or column selection

Often times one might make a selection and need to clear a selection before continuing.

To clear column selection click in the top right section as shown by the arrow



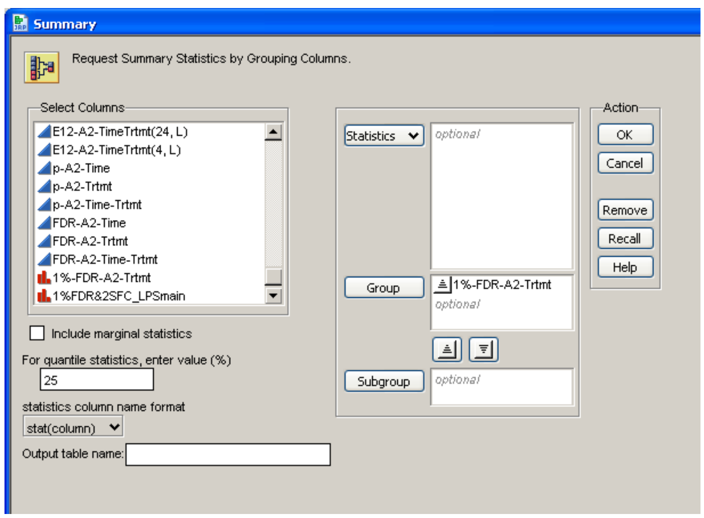
To clear row selection click in the bottom left section as shown by the arrow



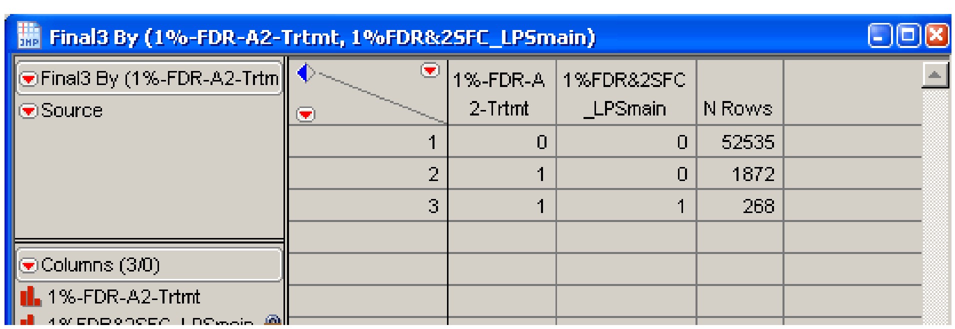
D. Table functions

1. Summary

   Summary is helpful categorize data by values in a column of choice.
   To used summary select **Tables→ Summary**

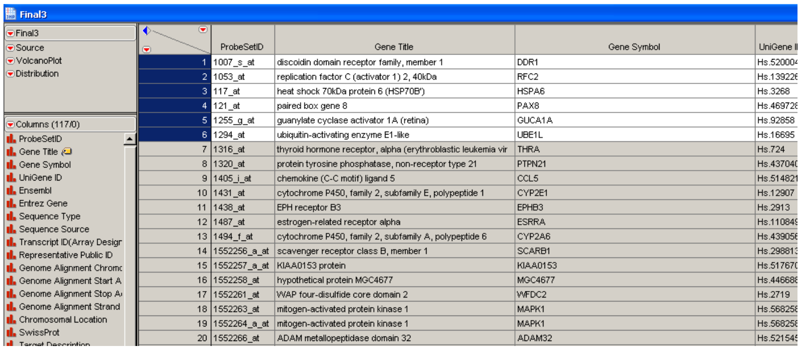


JMP will create a new data table containing the summary. Here is an example of a data table create by *Summary*



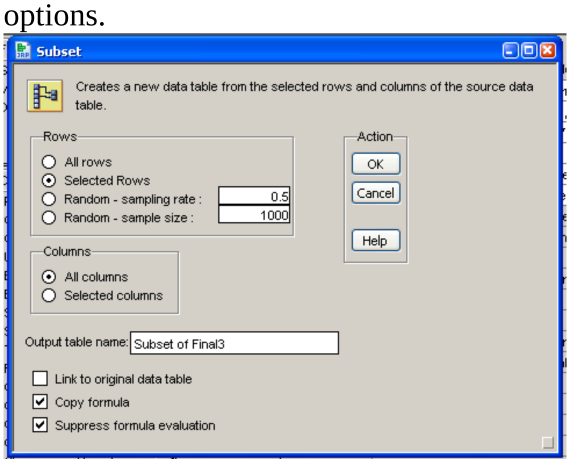
2. Subset
   Often there is a need to create a subset in order to focus on observations of interest. Subset allows you to reduce the number of both rows and/or columns To create a subset, first select the rows and/or columns of interest.

   Here we have a datatable with 20 rows and only rows are highlight in order to make a subset of six rows.

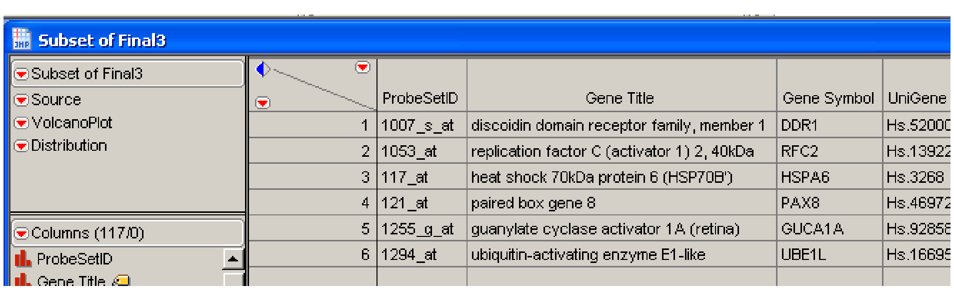


   Then select **Tables → Subset**

   The dialog box will appear and allow you to choose some row and column options.



   A new data table will be created and named Subset of the original table name as seen below



3. Join by matching columns
   Joining two tables on common columns is a useful tool and sets JMP apart from other software packages. For example we have a data table that contains data, in this case signal data, along with a probe set id and we would like to join this data with a table that contains gene titles and symbols. In this example we have 14 observations with signal data and only 10 gene titles.

   To use join first open the two data tables that need to be joined in JMP
   Here is the first; notice it has 14 rows and one column is ProbeSetID.

SignalData

	ProbeSetID	SG-L04R1	SG-L04R2	SG-L04R3	SG-L04R4	SG-L08R1	SG-L08R2
1	1007_s_at	230.847977	158.347565	177.097107	221.95929	190.64389	241.843597
2	1053_at	141.296432	158.856827	139.544128	143.970215	190.105103	131.581039
3	117_at	28.067181	21.634838	21.537586	16.763334	19.256332	39.465336
4	121_at	102.485374	120.561058	93.606659	117.068687	109.314407	115.974098
5	1255_g_at	4.508531	4.189002	4.065492	4.48659	13.822474	2.029952
6	1294_at	100.306725	126.132622	119.335831	112.798531	133.400024	82.770081
7	1316_at	12.800097	9.48455	25.296856	16.294512	14.31593	17.05883
8	1320_at	20.061094	3.366832	23.391787	11.244589	24.929869	4.172204
9	1405_i_at	207.233765	182.633194	175.604263	195.974594	185.494675	282.870056
10	1431_at	1.676353	12.445448	14.573278	4.842243	7.393765	11.765902
11	1438_at	6.601536	1.946463	5.648182	4.829084	4.619644	3.125274
12	1487_at	134.054062	159.394501	131.79863	142.263275	142.69101	180.224152
13	1494_f_at	13.029743	17.566202	19.093124	24.810963	23.373333	15.359431
14	1552256_a_at	82.548004	81.647438	75.304718	80.705315	99.56601	95.732056

Columns (7/0): ProbeSetID, SG-L04R1, SG-L04R2, SG-L04R3, SG-L04R4, SG-L08R1, SG-L08R2
All rows 14

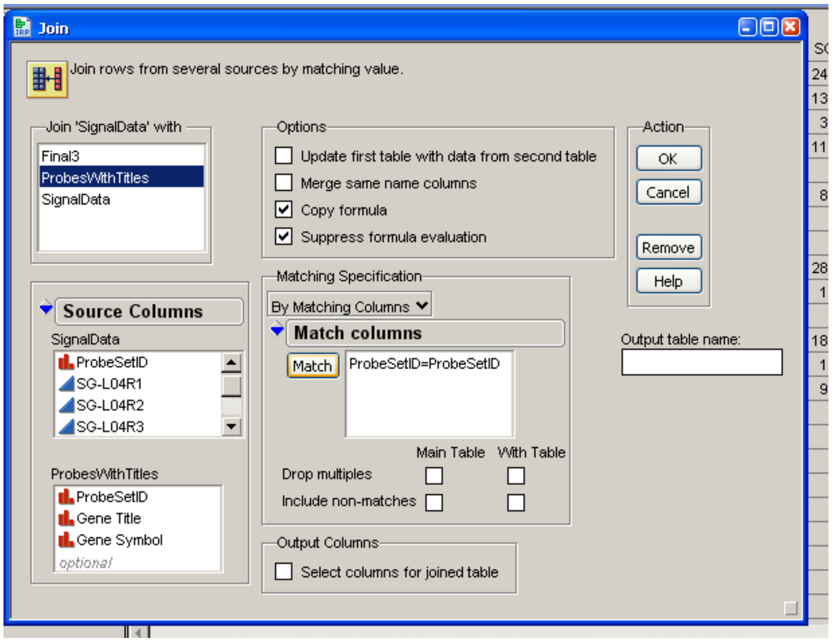
   Here is the second table with the information that needs to be joined. This table contains a column named ProbeSetID.

ProbesWithTitles

	ProbeSetID	Gene Title	Gene Symbol
1	1007_s_at	discoidin domain receptor family, member 1	DDR1
2	1053_at	replication factor C (activator 1) 2, 40kDa	RFC2
3	121_at	paired box gene 8	PAX8
4	1255_g_at	guanylate cyclase activator 1A (retina)	GUCA1A
5	1294_at	ubiquitin-activating enzyme E1-like	UBE1L
6	1320_at	protein tyrosine phosphatase, non-receptor type 21	PTPN21
7	1405_i_at	chemokine (C-C motif) ligand 5	CCL5
8	1438_at	EPH receptor B3	EPHB3
9	1487_at	estrogen-related receptor alpha	ESRRA
10	1494_f_at	cytochrome P450, family 2, subfamily A, polypeptide 6	CYP2A6

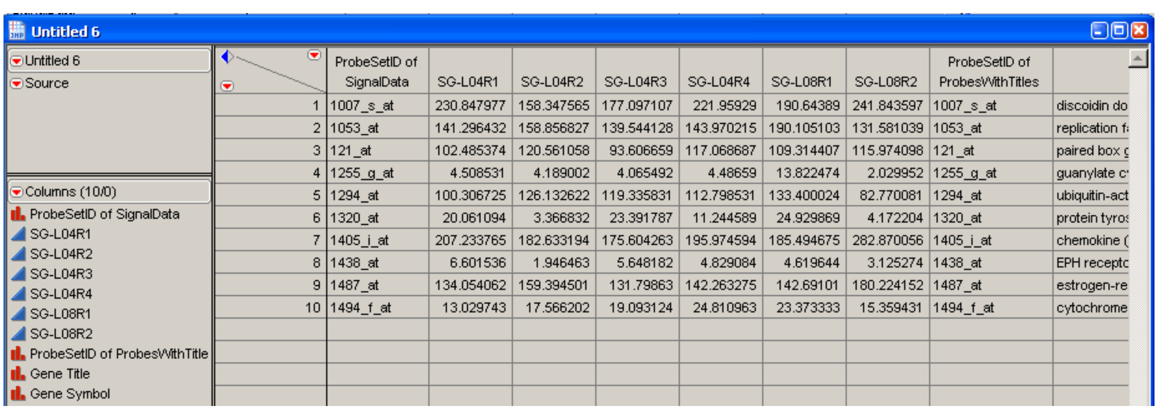
Columns (3/0): ProbeSetID, Gene Title, Gene Symbol
All rows 10

Once both tables are open select **Tables→ Join**.



There are several selections that must be made at this point. First choose the data table that you want to join with original data table. Then under the Matching Specification section choose By Matching Columns, as shown above. The last step is to click on ProbeSetID for both tables under Source Columns then click on Match.

The resulting data table will include those rows that shared a value in the joining column. Notice the columns from both tables are not in the new data table.



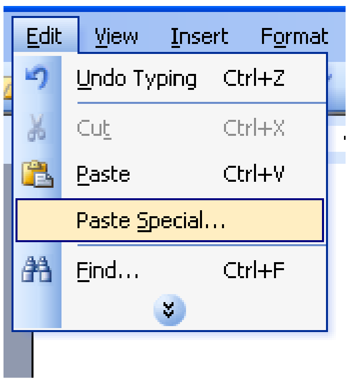
	ProbeSetID of SignalData	SG-L04R1	SG-L04R2	SG-L04R3	SG-L04R4	SG-L08R1	SG-L08R2	ProbeSetID of ProbesWithTitles	
1	1007_s_at	230.847977	158.347565	177.097107	221.95929	190.64389	241.843597	1007_s_at	discoidin do
2	1053_at	141.296432	158.856827	139.544128	143.970215	190.105103	131.581039	1053_at	replication f
3	121_at	102.485374	120.561058	93.606659	117.068687	109.314407	115.974098	121_at	paired box g
4	1255_g_at	4.508531	4.189002	4.065492	4.48659	13.822474	2.029952	1255_g_at	guanylate c
5	1294_at	100.306725	126.132622	119.335831	112.798531	133.400024	82.770081	1294_at	ubiquitin-act
6	1320_at	20.061094	3.366832	23.391787	11.244589	24.929869	4.172204	1320_at	protein tyros
7	1405_i_at	207.233765	182.633194	175.604263	195.974594	185.494675	282.870056	1405_i_at	chemokine (
8	1438_at	6.601536	1.946463	5.648182	4.829084	4.619644	3.125274	1438_at	EPH recepto
9	1487_at	134.054062	159.394501	131.79863	142.263275	142.69101	180.224152	1487_at	estrogen-re
10	1494_f_at	13.029743	17.566202	19.093124	24.810963	23.373333	15.359431	1494_f_at	cytochrome

E. **Saving and exporting graphics**

Once you are finished analyzing data in JMP there are a several options for saving your results. Here are two.

1. As bitmap

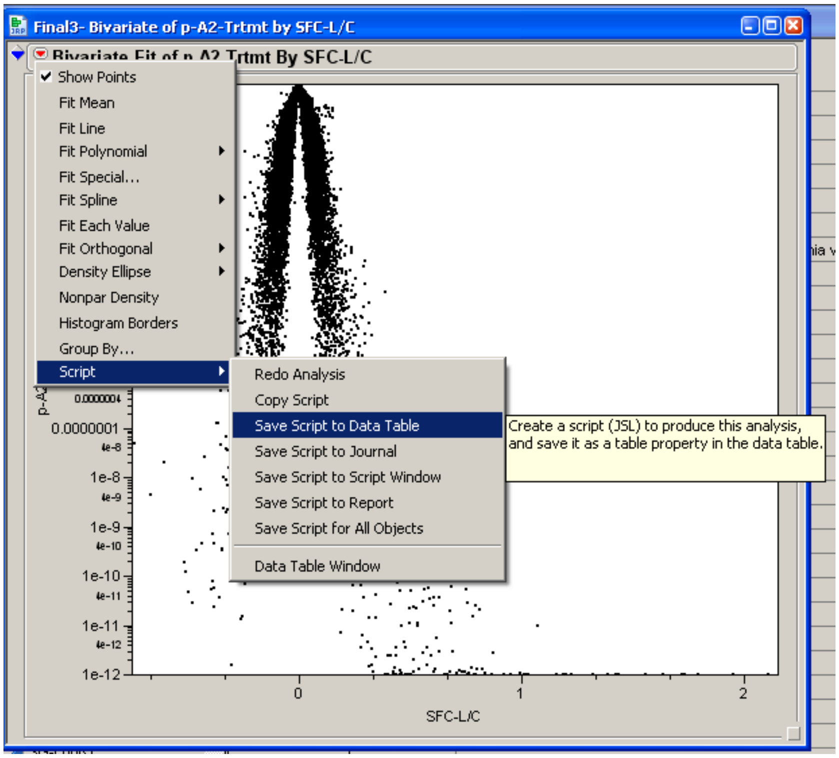
   Bitmap is form that seems to be standard enough to be read with no problem on both Mac and PC.
   To save a file as a bitmap first create the plot or graph then select **Edit → Copy.** Open the a Word or PowerPoint document and then paste by using *Special Paste*



2. Saving script to table
   Any visualization tool output in JMP can be stored in the data table right along with the data.

   To save output with your data, JMP save the data as a script. To do this click on the red triangle in the top right hand corner of the output window, as seen here 

   Then select **Script→ Save Script to Data Table**


Final3- Bivariate of p-A2-Trtmt by SFC-L/C
Show Points
Fit Mean
Fit Line
Fit Polynomial
Fit Special...
Fit Spline
Fit Each Value
Fit Orthogonal
Density Ellipse
Nonpar Density
Histogram Borders
Group By...
Script
Redo Analysis
Copy Script
Save Script to Data Table
Save Script to Journal
Save Script to Script Window
Save Script to Report
Save Script for All Objects
Data Table Window
Create a script (JSL) to produce this analysis, and save it as a table property in the data table.
0.0000001
1e-8
1e-9
1e-10
1e-11
1e-12
0
1
2
SFC-L/C

Chapter 3:
Analyzing microarray data with the Toolbox

Depending on what type of chips you are using, there are multiple formats that your data can be in. If you are using Affymetrix then the MSCL toolbox provides three ways to import your data. The *GetAnalysisNames* and the *FetchData* scripts are provided and will fetch the data from the GCOS database locally. If you are using GCOS but do not have the database set up locally, then you can download the PivotData.txt from GCOS or from EC and use JMP's *Text Import Preview* dialog to import the data. The data should be in the format with samples as columns and genes or probe sets as rows.

Importing data, 4 options:

A. Import data into JMP using text file
B. Import data into JMP using PivotData.txt generated from Affymetrix GCOS
C. Fetch data into JMP using a PivotData.txt generated from Affymetrix EC
D. Fetch data into JMP from GCOS database using MSCL toolbox scripts

Master File:

File containing sample information about the experimental design along with original file names and ShortNames used for easier data manipulation. This file contains experiment samples as rows and information about each sample as columns. This file should contain a column indicating the type of treatment or tissue type that will be used for the differential expression testing for each chip.

Final File:

File containing actual intensity data for each probe set. File format is samples as columns and genes as rows. Sample names in this file are a prefix of type of data (SG for signal, AD for average difference) and ShortNames that link to MasterFile.

An outline for processing and analyzing microarray data

1. Fetch Data and/or import expression data to create Final table and Master File

   Affymetrix GCOS:

   a. Run parse Affy pivot table from GCOS
   b. Run recode Affy pivot table into Final table from GCOS

   Affymetrix Expression Console (EC)

   a. Run parse Affy pivot table from EC
   b. Run recode Affy pivot table into Final table from EC
2. Normalize/transform data if necessary
3. Compute principal components analysis (PCA)
   a. create master file if necessary or join PCA results table with master file
   b. inspect PC plots for outliers, experiment success
4. Set up experimental design in master file adding categorical variables
5. Choose and run appropriate statistical tests to generate p-value and False Discovery Rate (FDR) correction
6. Create gene list indicator columns
   a. add filter criteria if applicable
      1. p-value cutoff
      2. fold change cutoff
      3. present/absent call cutoff

7. Visualize data using graphical features
8. Thematic searches, gene list discovery, pathway discovery i.e. GO-SCAN, EASE, IPA, GSEA, GO-Miner

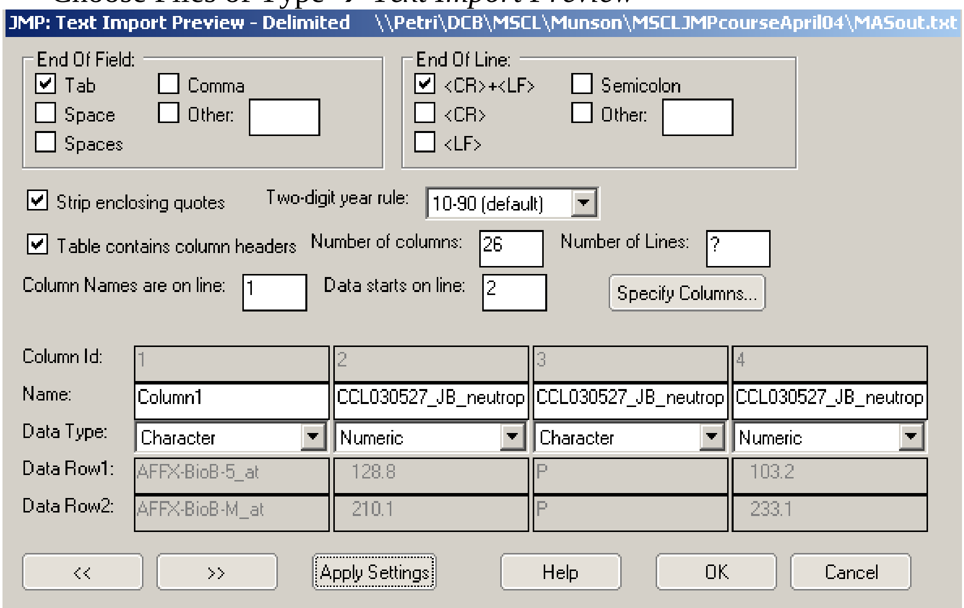
11. Validate data i.e. RTPCR

Importing your data

There are 4 ways to get your data into JMP as described on the previous page.

A. Import data into JMP using text file
B. Import data into JMP using *PivotData.txt* generated from Affymetrix GCOS
C. Import data in JMP using *PivotData.txt* generated from Affymetrix EC
D. Fetch data into JMP from NIHGCOS database using MSCLToolbox

A. Importing a text file into JMP, use *Text Import Preview* under File tab in JMP
File→Open
Choose Files of Type → *Text Import Preview*



B. Using pivot table from Affymetrix GCOS

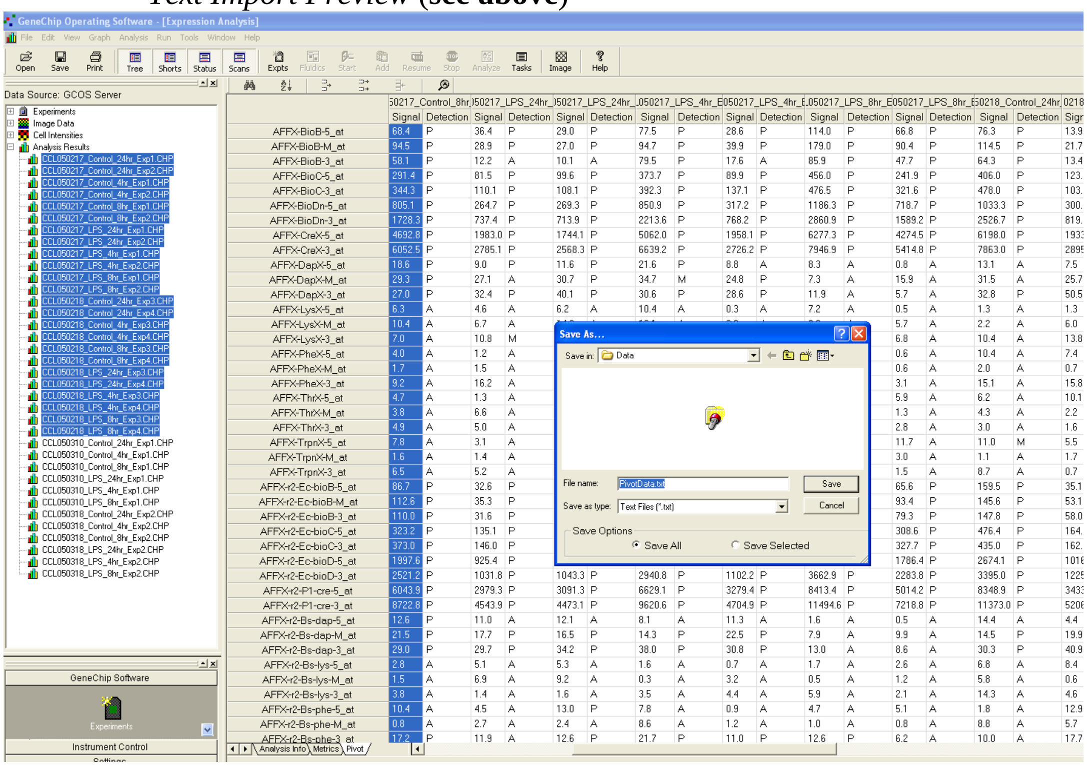
To create the pivot table in Affymetrix GCOS, the data should first be loaded into GCOS. To do that:

Set options in GCOS to display Signal and Detection data:

1. Go to Analysis tab → Select options
2. Click Pivot tab
3. Select: Signal, Detection and Avg Diff (Average Difference is pre MAS5.0, Signal is for MAS5.0 and present, Detection is Present/Absent call)
4. Click OK

Export data from GCOS as a pivot table

1. Selecting Analysis Results tab
2. Select the chips that you want to export
3. Right click over selected items and click open
4. Pivot table will be created in the current session
5. Go to File menu and click Save
6. This will give you a PivotData.txt file that can now be read into JMP using *Text Import Preview* (**see above**)



C. Using Pivot table data form Affymetrix EC.

To create the pivot table using Affymetrix EC, .CEL files are required. To do that:

From within Affymetrix EC:

First set library path and download libraries if you don't have Affy libraries

1. Go to Edit tab→ then browse to or create a folder for the library files.
2. Then fill the selected folder with library files or if you do not have the library files go to File tab→ select Download Library Files (this directs you to Affymetrix website login required)

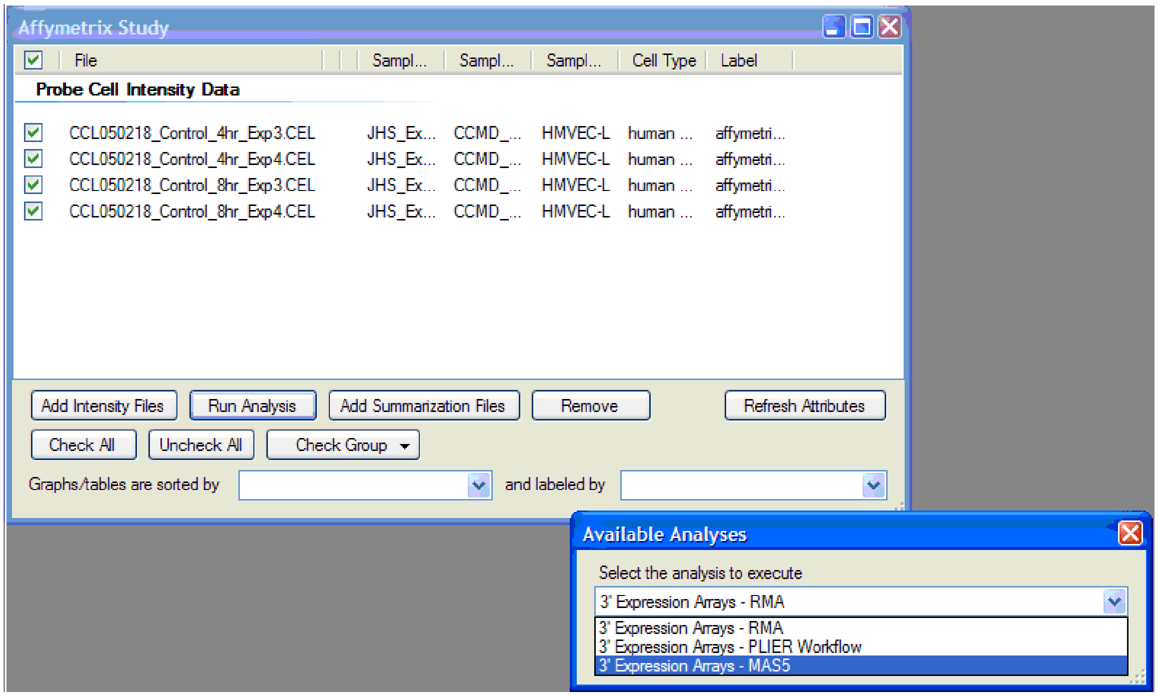
Set options in EC to display Signal and Detection data:

1. Go to Edit tab→ Probe Level Summarization Report Options
2. check the boxes Signal and Detection
3. Click OK

Reading .CEL files into EC

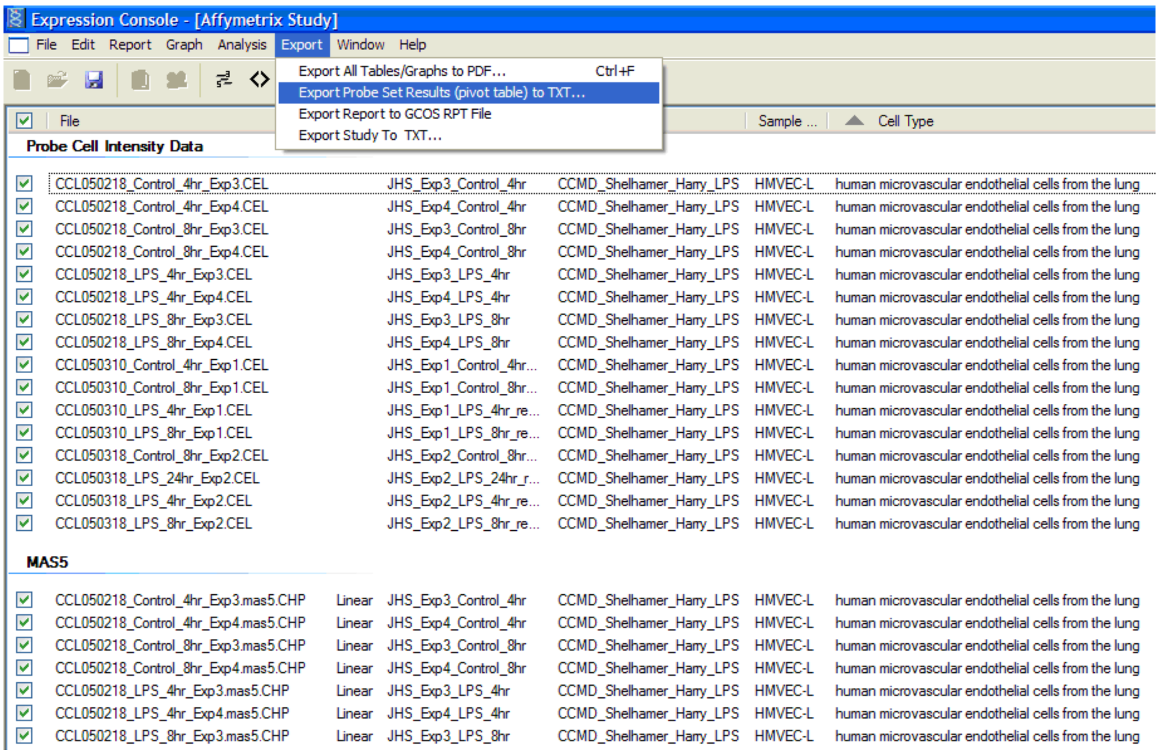
1. Go to File tab→ select New study
2. In the window that pops up select Add Intensity Files button and browse to your .CEL files.
3. Select your files (to select more than one use shift click) and click on Open button
4. Once the filenames appear in the window make sure all files are have a check in the box preceding the filename
5. Select Run Analysis and click on arrow to select 3' Expression Arrays- MAS5 then click OK
6. Choose a suffix or for no suffix just click OK

** for other information on EC, download EC notes from the MSCL toolbox website

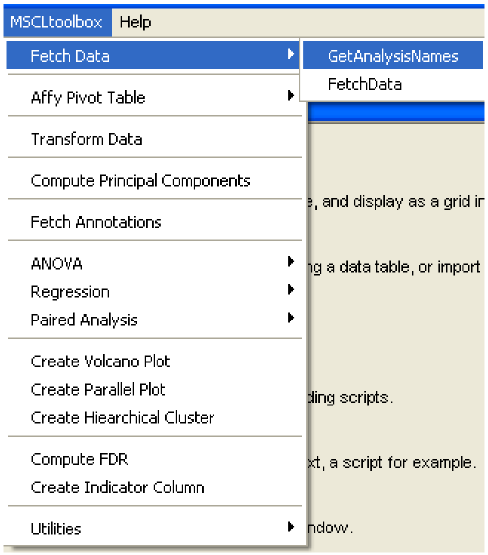


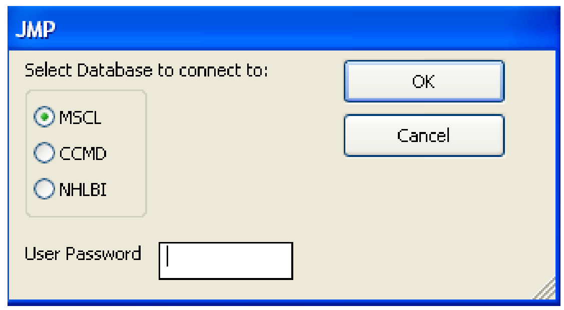
Export data from EC as a pivot table

1. Select the files to be exported by checking the box in front of the .CHP filenames.
2. Select the Export tab → Export Probe Set Results (pivot table) to TXT
3. Name the file PivotData.txt and click Save
4. Pivot table will be created in C:\Program Files\Affymetrix\ExpressionConsole folder
5. This will give you a PivotData.txt file that can now be read into JMP using *Text Import Preview* (**see above**)



D. Select *FetchData* → *GetAnalysisNames* from MSCLToolbox on toolbar and enter password to database, this will give you a list of all experiments in that database. You may now proceed to the next section to learn how to create Master File and Final File





Results are a list of all experiments in the database. This table will be used to create your MasterFile and your Final data table. Proceed to next section for instructions

Creating the Master File and Final data table

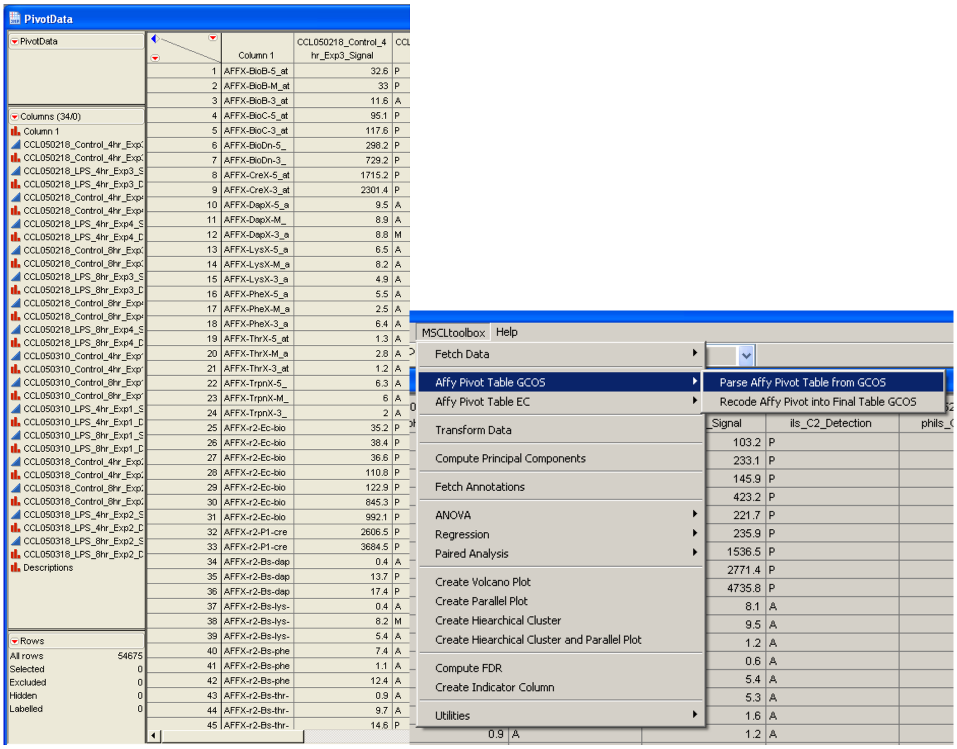
There are four ways to create MasterFile and Final data tables:

If you are not getting data from Affy GCOS, use A.
If you have data loaded into GCOS, use B.
If you have data loaded into EC, use C
If your data are stored in NIHGCOSA server, use D, below.

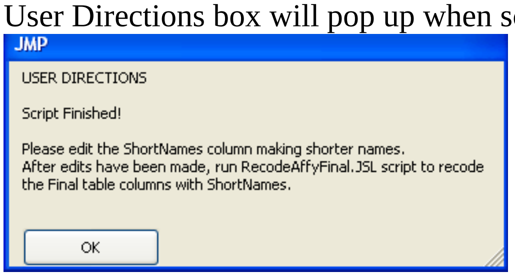
A. If you are not using Affymetrix and do no have a pivot table or are not fetching your data from GCOS, then you can create your master file from running the *ComputePCA* script (proceed to chapter 3 for data transformation and chapter 4 for principal components analysis)

B. If you are using a pivot table downloaded from GCOS then follow along here:

After pivot table is open in JMP, select *Affy Pivot Table GCOS*→ *Parse Affy Pivot Table from GCOS* from MSCL Toolbox Menu

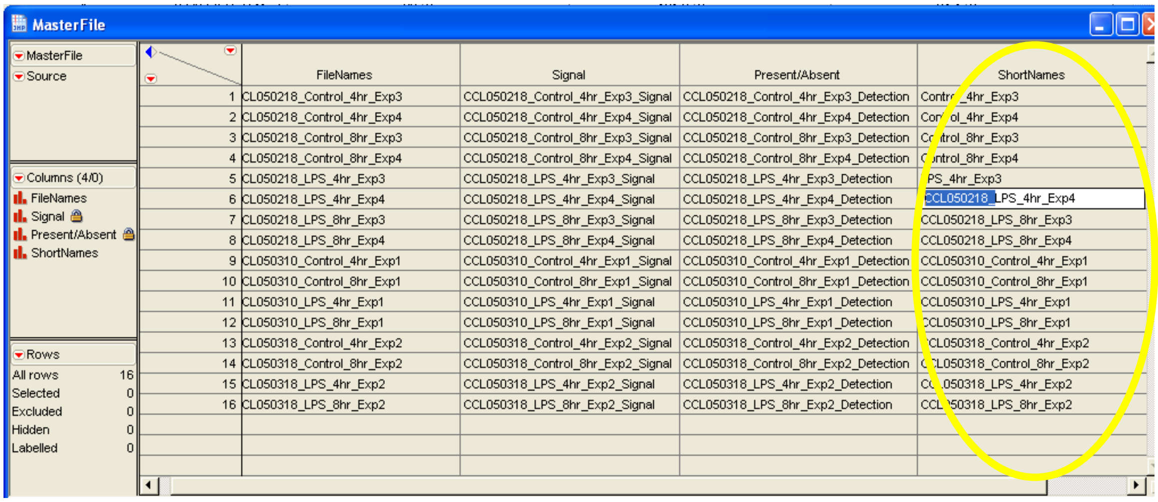


Script will look for Signal and Detection columns in pivot table.
Result of running script will be a MasterFile containing file name stems, Signal column names, Detection column names and ShortNames
User Directions box will pop up when script is finished.

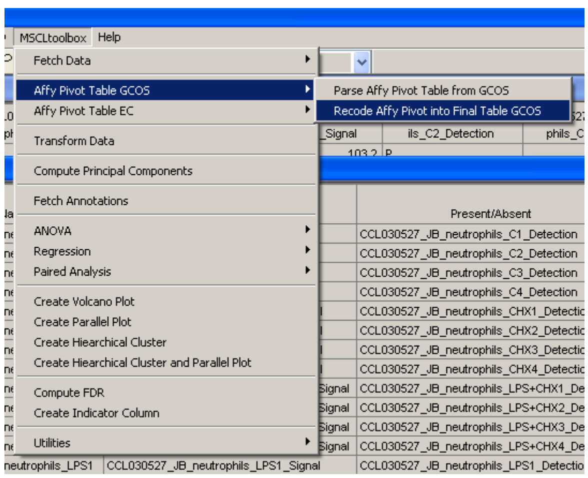


After MasterFile table is generated from running *Parse Affy Pivot Table from GCOS*, ShortNames column needs to be edited where typically shorter names signifying experiment treatments are created.

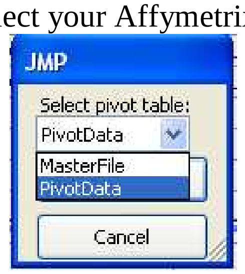
Edit ShortNames field in MasterFile output



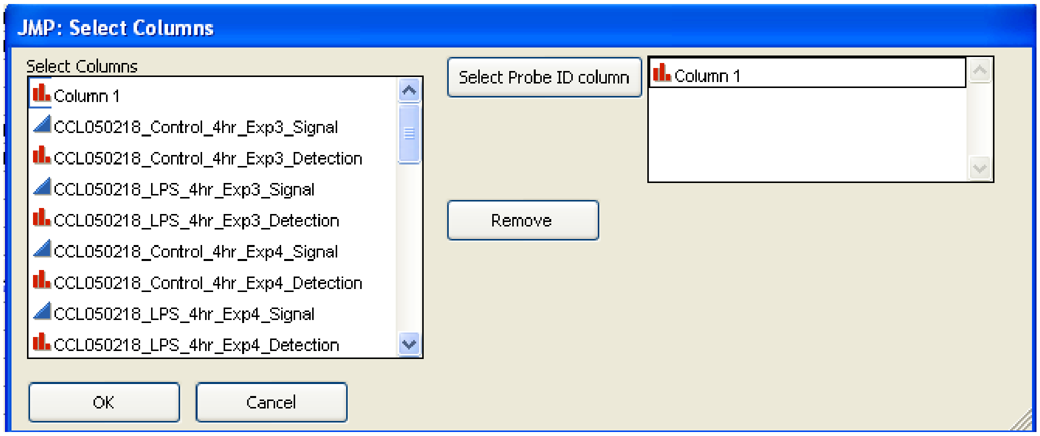
Select *Affy Pivot Table GCOS* → *Recode Affy Pivot into Final Table GCOS*



After script is selected, user dialog box will pop up shown below.
Select your Affymetrix pivot data Table



Select probe id column, usually the first one.

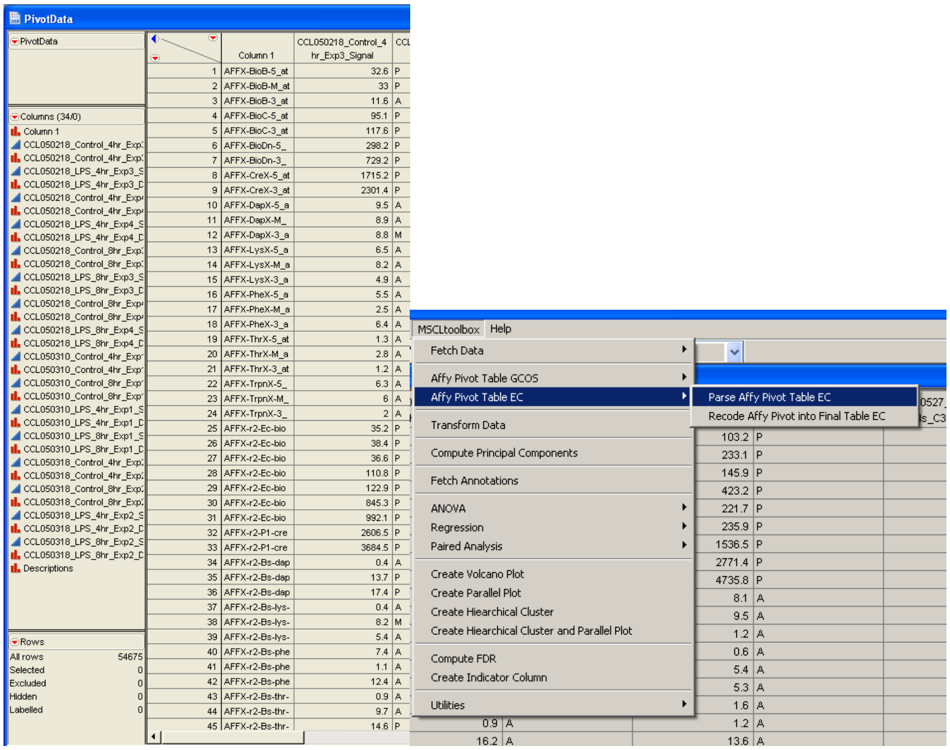


A Final table will be created from your Affymetrix pivot data, with prefixes appended: "SG-" for Signal columns ("AD-" for Avg Diff columns) and "PA-" appended for Detection columns. The script will also change Detection columns into numeric where "P" is changed to 1, "M" is changed to 0.5 and "A" is changed to 0.

Now you have a MasterFile and a Final table that are linked by ShortNames. You are now ready to begin Data transformation and Analysis, proceed to chapter 3

C. If you are using a pivot table downloaded from EC then follow along here:

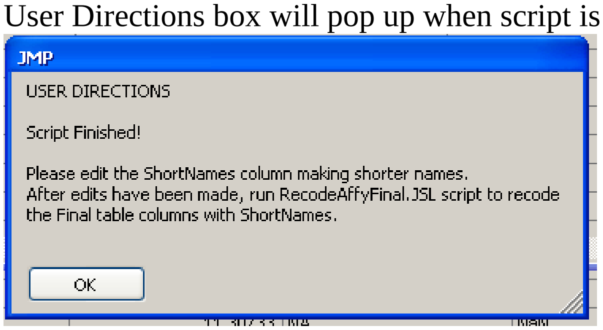
After pivot table is open in JMP, select *Affy Pivot Table EC* → *Parse Affy Pivot Table EC* from MSCL Toolbox Menu



Script will look for Signal and Detection columns in pivot table.

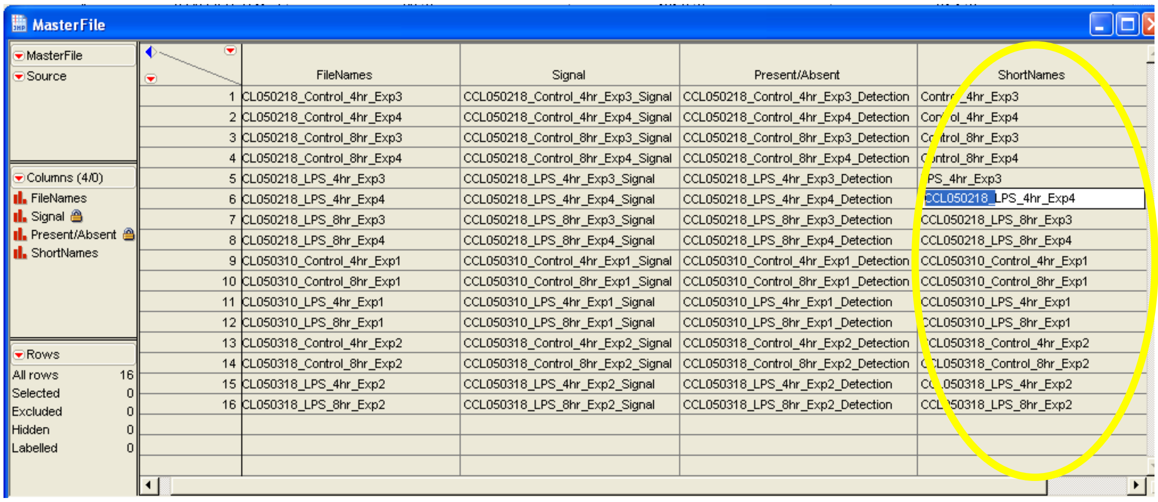
Result of running script will be a MasterFile containing file name stems, Signal column names, Detection column names and ShortNames

User Directions box will pop up when script is finished.

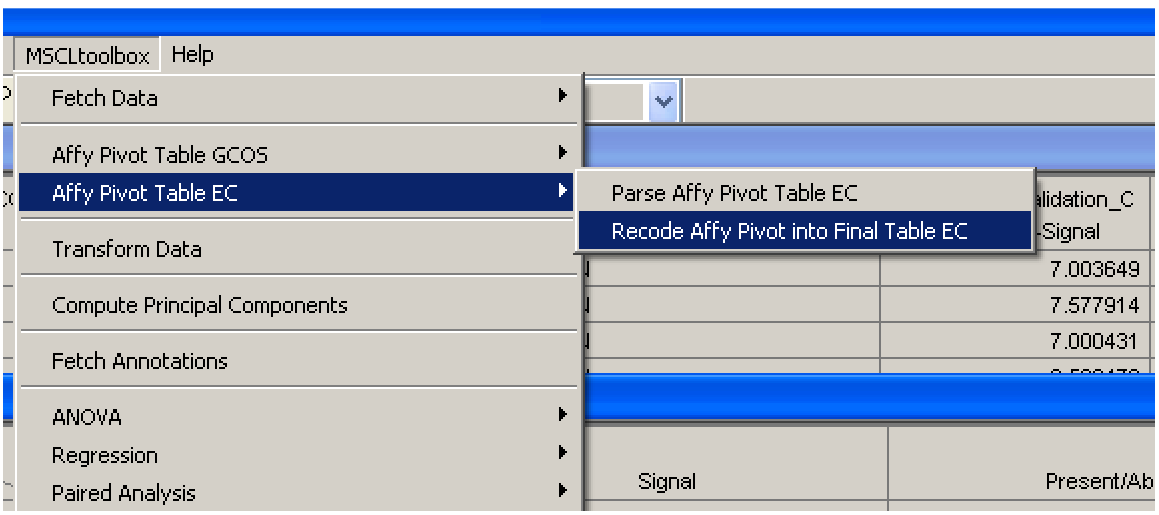


After MasterFile table is generated from running *Parse Affy Pivot TableEC*, ShortNames column needs to be edited where typically shorter names signifying experiment treatments are created.

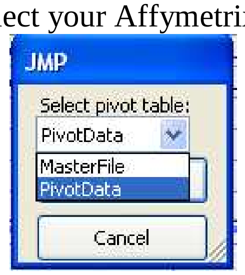
Edit ShortNames field in MasterFile output



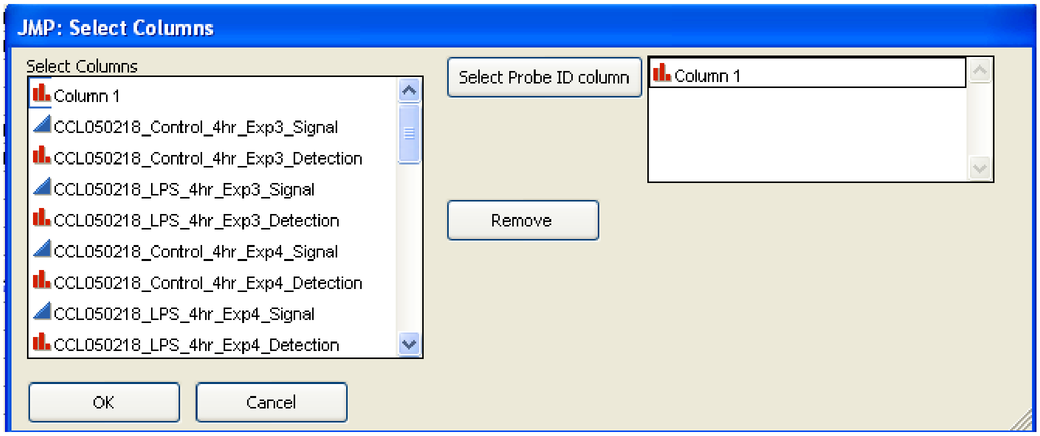
Select *Affy Pivot Table EC* → *Recode Affy Pivot into Final TableEC*



After script is selected, user dialog box will pop up shown below.
Select your Affymetrix pivot data Table



Select probe id column, usually the first one.

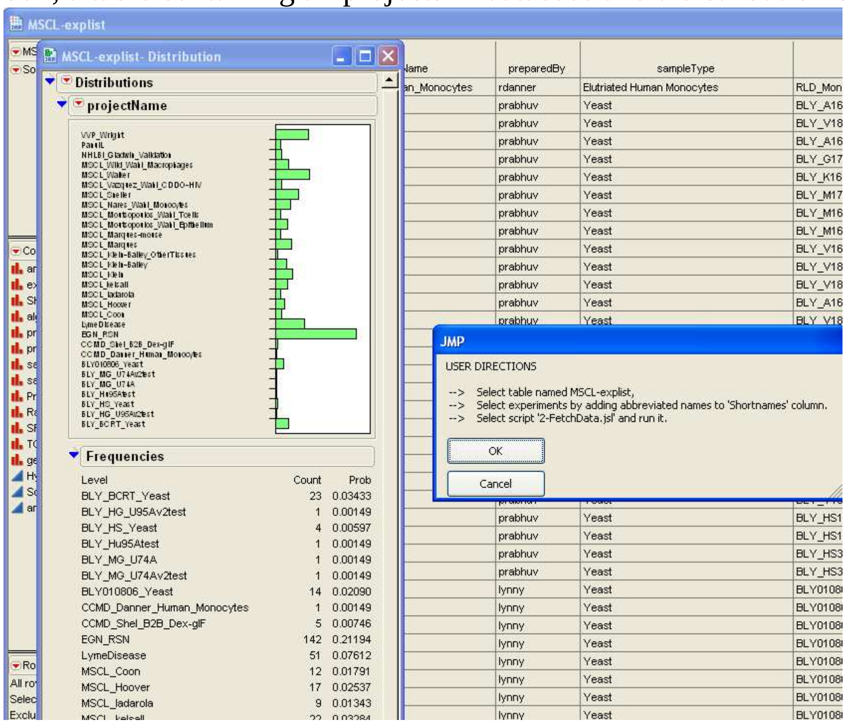


A Final table will be created from your Affymetrix pivot data, with prefixes appended: "SG-" for Signal columns and "PA-" appended for Detection columns. The script will also change Detection columns into numeric where "P" is changed to 1, "M" is changed to 0.5 and "A" is changed to 0.

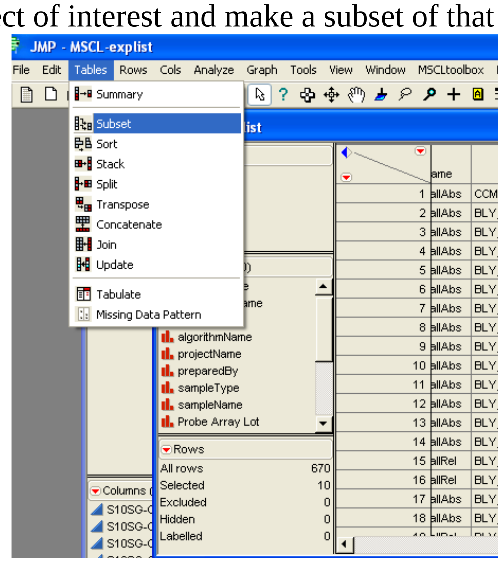
Now you have a MasterFile and a Final table that are linked by ShortNames. You are now ready to begin Data transformation and Analysis, proceed to chapter 3

D. If you downloaded your experiment data using the *GetAnalysisNames* script, then follow along here:

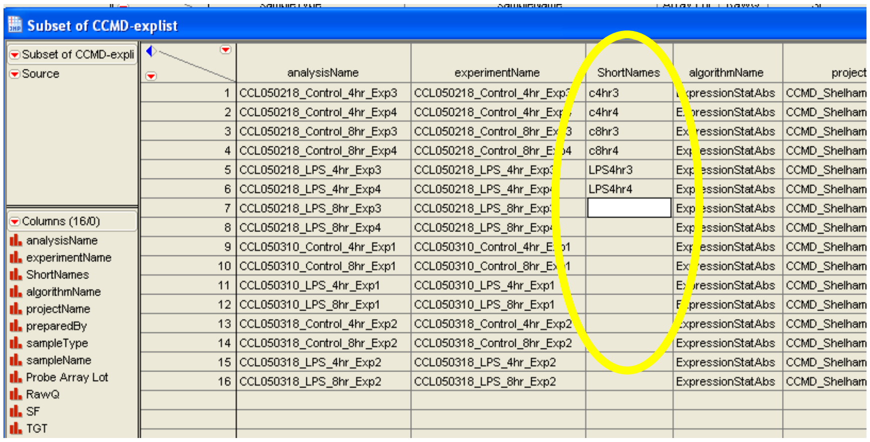
When *GetAnalysisNames* script is finished running, you will get a Users Directions box, a table containing all projects in database and a distribution of all projects.



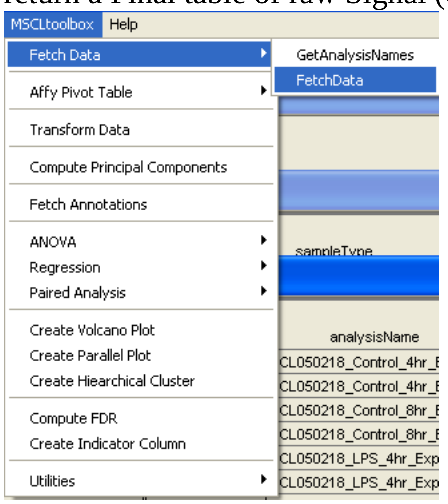
Select project of interest and make a subset of that project:



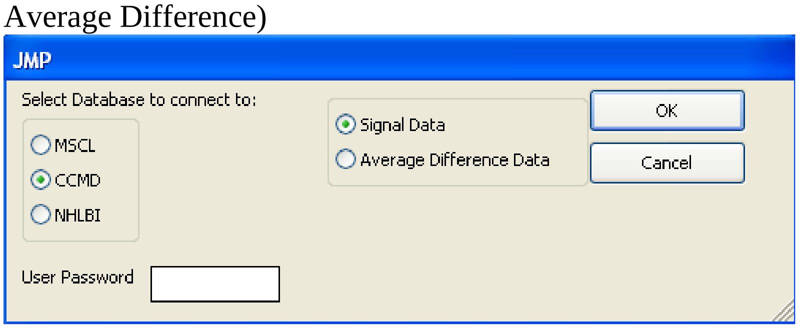
Begin editing ShortNames field . This file can now serve as your MasterFile, save as MasterFile.



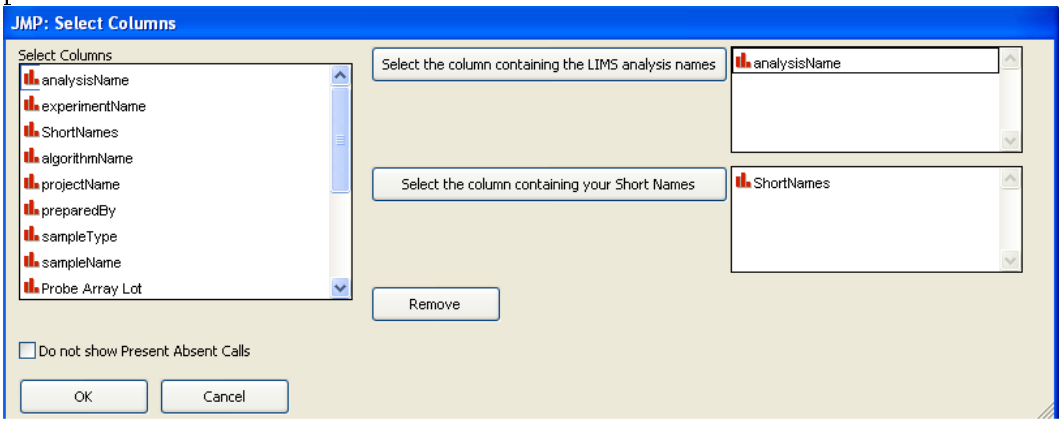
Run *Fetch Data* → *FetchData* script from MSCLToolbox menu
This script will fetch the experiments with ShortNames entries in the MasterFile and will return a Final table of raw Signal (or Avg Diff) values and a probe id column.



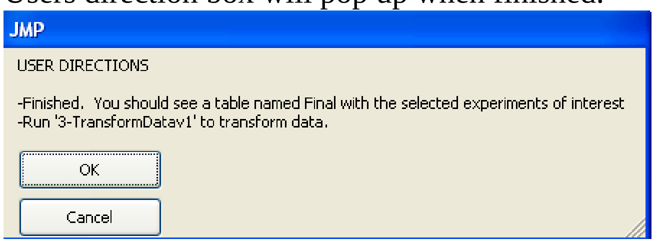
Select which database to use and which type of data normalization is used (Signal, Average Difference)



Select analysisName column and ShortNames column.
If you want Present Absent calls printed, then leave dialog box unchecked, default is to print them.



Users direction box will pop up when finished.

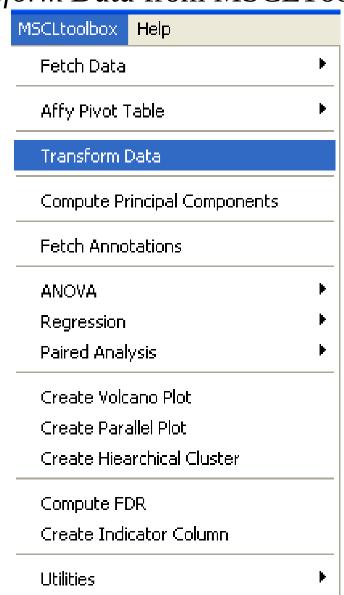


Result of script is a Final table. You are now ready to begin data transformation. Move on to Chapter 3.

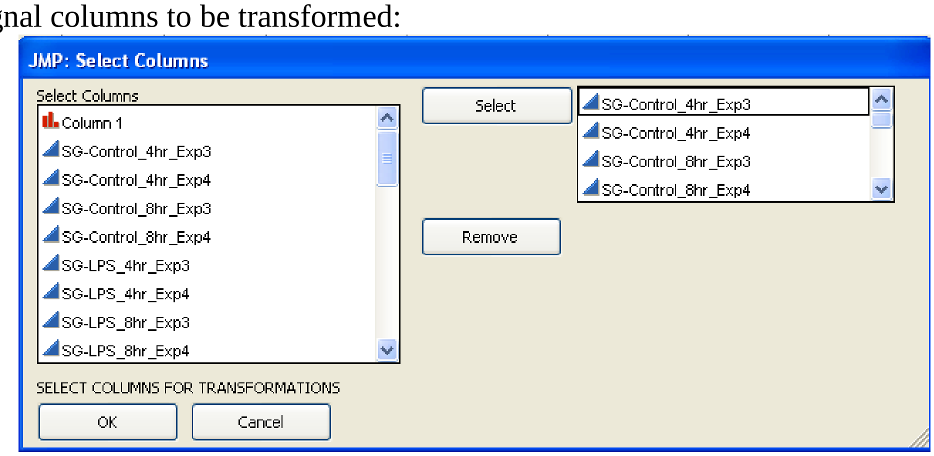
Chapter 4:
Data Transformation

There are a number of different data transformations that can be run from the MSCL toolbox. Depending on the microarray platform used, it is important to choose the appropriate transformation. There are 6 different transforms that can be chosen.

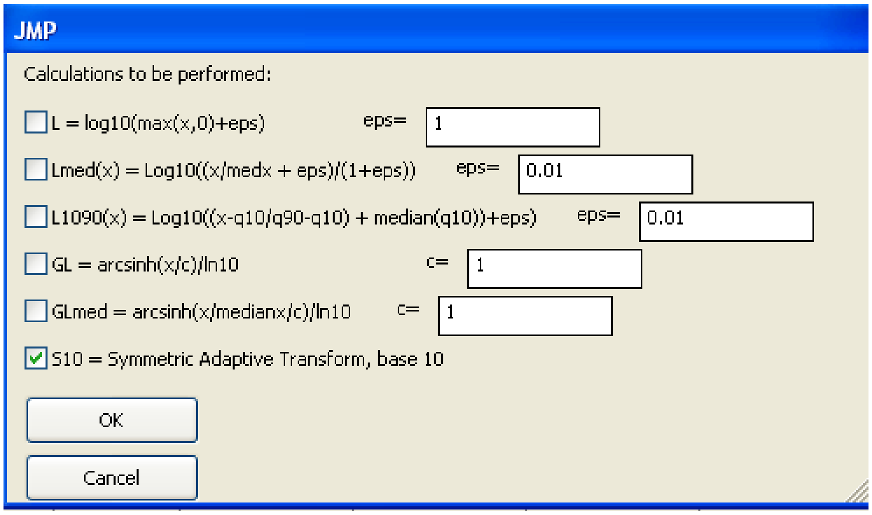
Run *Transform Data* from MSCLToolbox menu



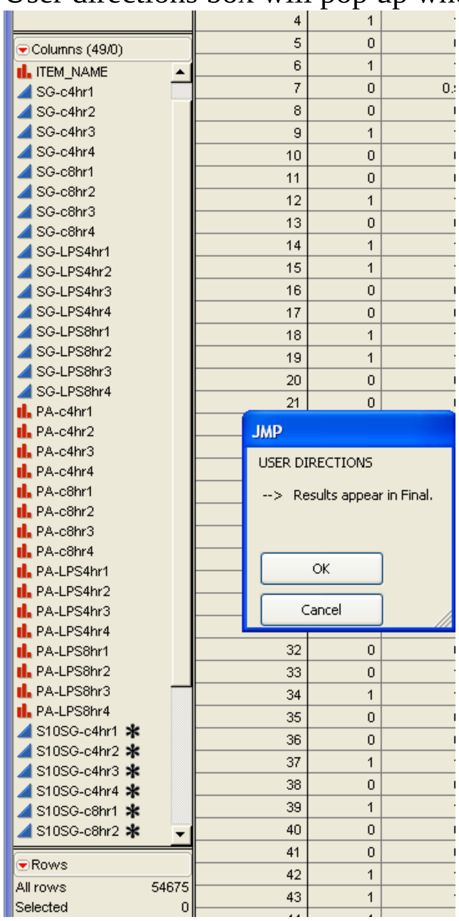
Select Signal columns to be transformed:



Choose transform of interest



Transform columns will be appended to end of Final data table with appropriate prefix appended to the column names (ie. S10 for Symmetric Adaptive Transform, base 10). User directions box will pop up when script is finished running.

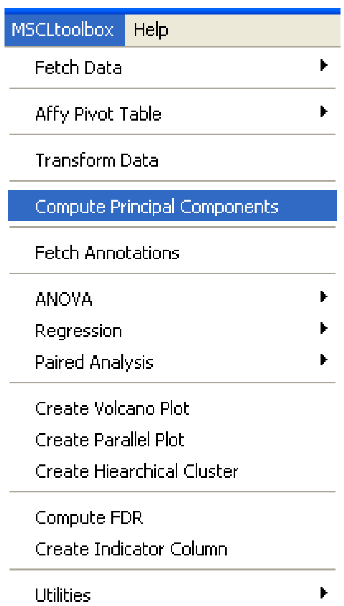


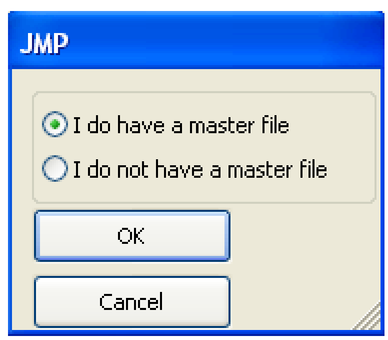
You are now ready to run the Principal Components Analysis, Chapter 4

Chapter 5:
Principal Components Analysis

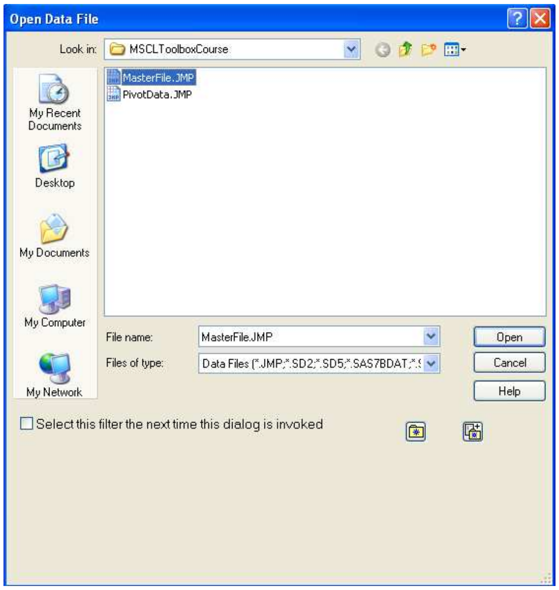
If you already created your MasterFile the *ComputePrincipalComponents* will ask you to open that data set so that the principal components may be joined into it. If you are not using Affymetrix as your platform but have your data in the appropriate format (i.e. columns as samples and rows as genes) then you may proceed with running the *ComputePrincipalComponents* script in order to create your MasterFile however you should have already transformed your dataset as was described in Chapter 3. You will want to run a PCA on transformed/normalized data. Choose *Compute Principal Components* in the MSCLToolbox menu bar.

Select whether you have a master file or not from the dialog box that pops up after running the script. If you do have one, open it. If you do not have one, choose "I do not have a master file" and click OK.

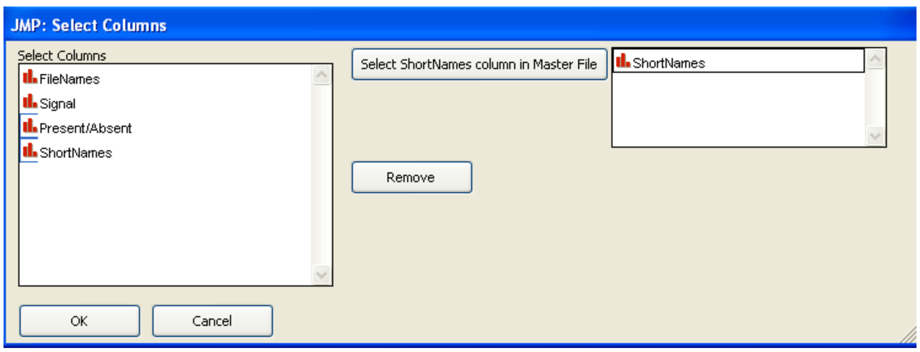




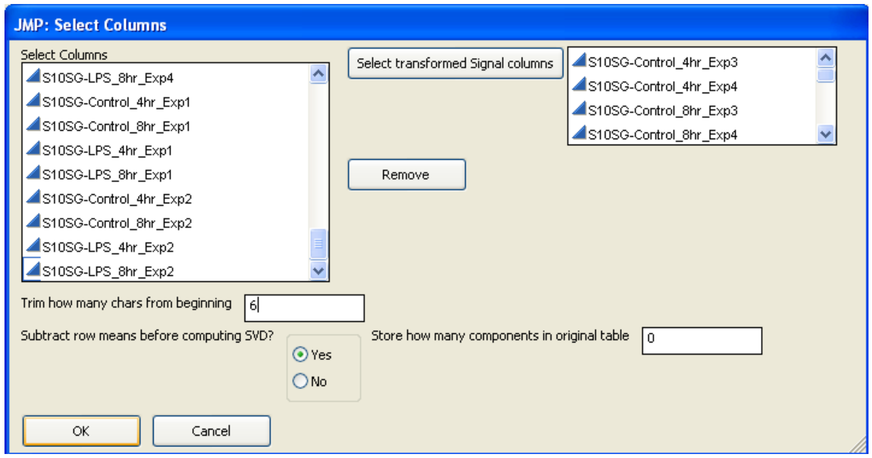
Open MasterFile associated to the Final data table if you have one.



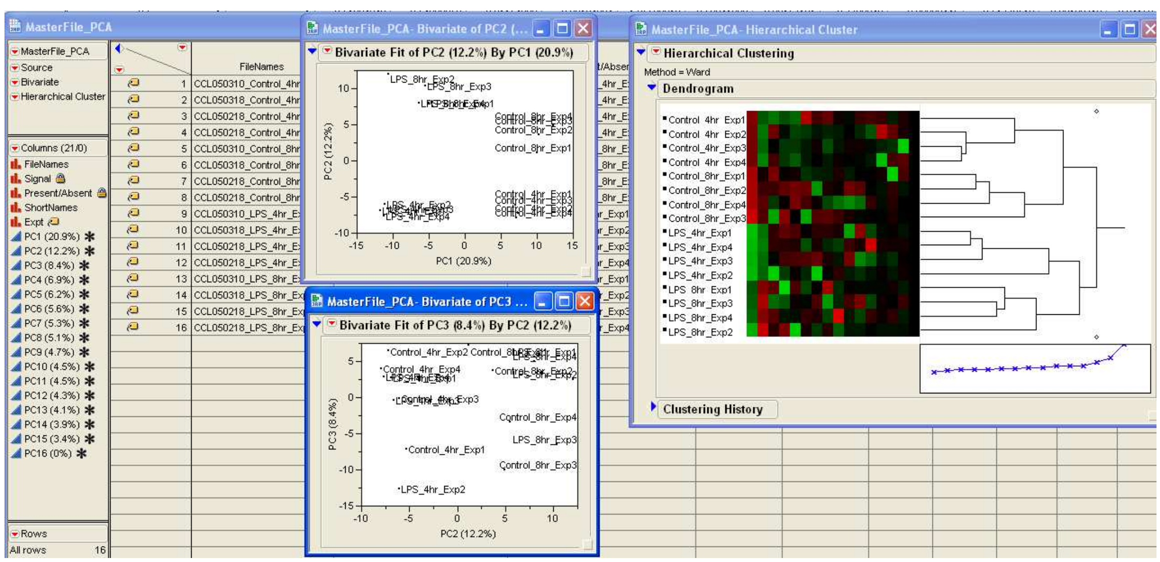
Select ShortNames column in Master File



Select Transformed signal columns for PCA computation and the correct number of characters to trim off of column names (usually however many characters the transform pre-pended).
**The ShortNames column in the Master File directly corresponds to the expression column names in the Final file. Trimming off the prefix characters from the expression column names should result with short names that match the rows in the Master File. See course notes Day 1 on the MSCL toolbox web page.



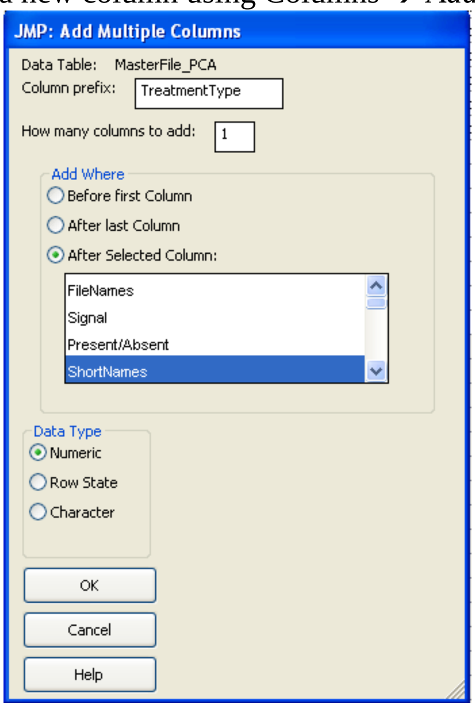
Output from script is a Master File (if you had one created already) with principal components joined or a PCAtable (if you did not have one created already which can now serve as your MasterFile), 2 bivariate plots plotting the first 3 principal components (PC2 vs PC1 and PC3 vs PC2) and a Hierarchical Cluster plot of all PC's (PC's as columns and samples as rows)



Chapter 6:
Experimental Design Setup

The experimental design is important for the statistical test used. The Master file should contain information about each sample such as clinical data, date of analysis, sample information, treatment information etc. In order for the statistical tests to run, the master file will also need information on each sample that will tell it to which treatment group each sample belongs to. This chapter shows an example of adding a new column to the master file and filling in the sample information/experimental design to be used by any of the statistical scripts.

Create a new column to add the experimental design information in the Master File. Create a new column using Columns → *Add Multiple Columns* or → *New Column* button

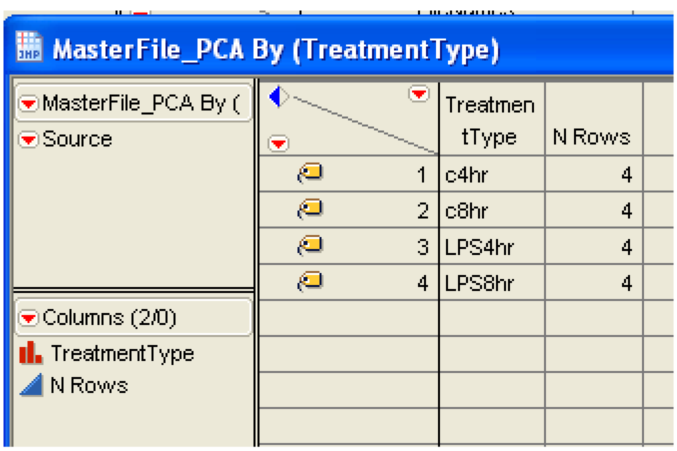


Fill in each row of table with the appropriate information for each sample. In this example, the first 4 rows correspond to the 4hr control samples, the next 4 rows correspond to the 8hr control samples, the following 4 rows correspond to the 4hr treated samples and the final 4 rows correspond to the 8hr treated samples. This particular example/experiment thus contains 4 groups with 4 replicates per group for a total of 16 samples/chips.

ShortNames	TreatmentType	
Control_4hr_Exp1	c4hr	Cor
Control_4hr_Exp2	c4hr	Cor
Control_4hr_Exp3	c4hr	Cor
Control_4hr_Exp4	c4hr	Cor
Control_8hr_Exp1	c8hr	Cor
Control_8hr_Exp2	c8hr	Cor
Control_8hr_Exp3	c8hr	Cor
Control_8hr_Exp4	c8hr	Cor
LPS_4hr_Exp1	LPS4hr	LPS
LPS_4hr_Exp2	LPS4hr	LPS
LPS_4hr_Exp3	LPS4hr	LPS
LPS_4hr_Exp4	LPS4hr	LPS
LPS_8hr_Exp1	LPS8hr	LPS
LPS_8hr_Exp2	LPS8hr	LPS
LPS_8hr_Exp3	LPS8hr	LPS
LPS_8hr_Exp4	LPS8hr	LPS

The above entered in experimental design of the data will generate a summary of 4 levels with 4 replicates for each level (see below).

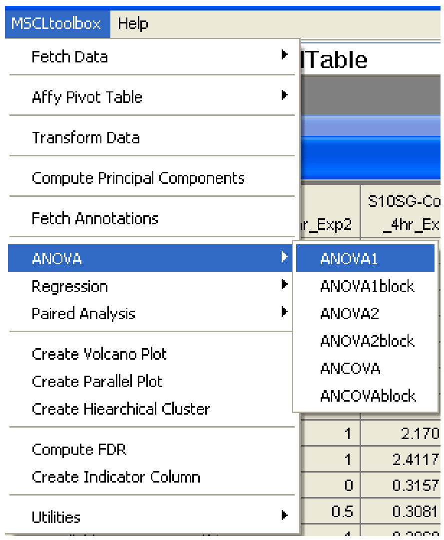
Executing a Summary on the new column from the Tables→ Summary tab shows this design:



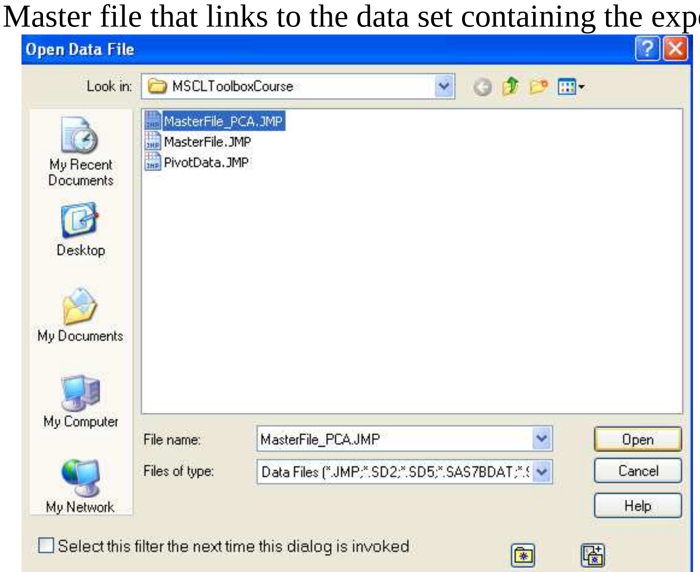
Chapter 7:
Statistical tests and Indicator columns

There are numerous statistical tests available in the toolbox including Analysis of Variance (ANOVA), Regression and Paired Analysis tests. Each of these tests depend on a Master File and its linkage to the Final file. The master file must contain the experimental design information appropriate to the statistical test being run.

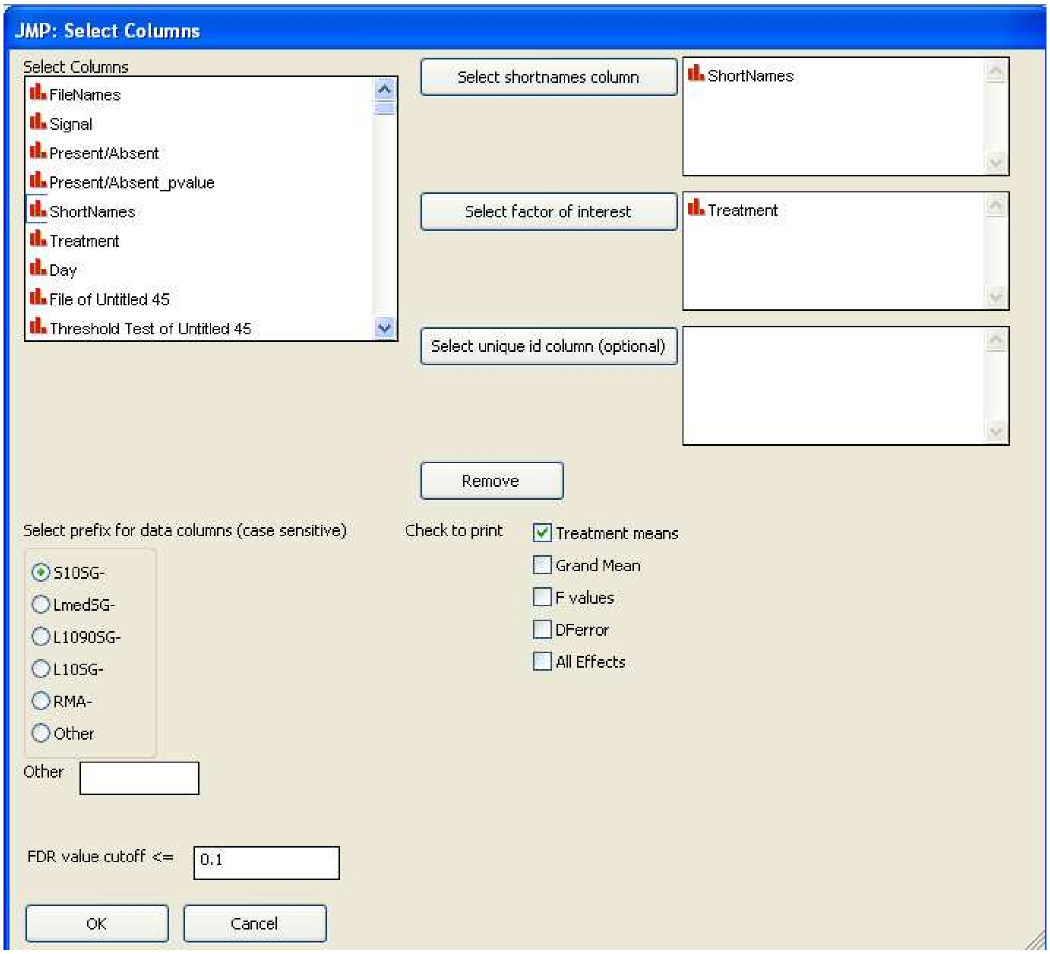
A. One Way ANOVA test
   Categorical column in Master file should contain more than 1 group for the one and more than 1 replicate for each group for the one way ANOVA test
   Choose *ANOVA* → *ANOVA1* on the toolbar menu



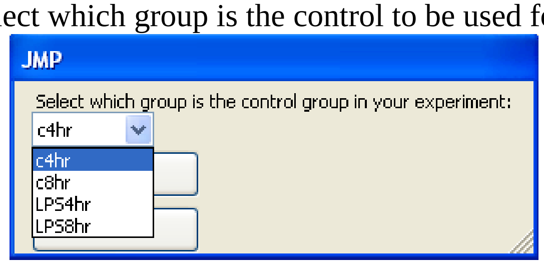
Open Master file that links to the data set containing the experimental design column(s):



Select ShortNames column and experimental design column containing experimental information.
Type in or select prefix that matches the expression column names in the Final data table.
Check to print out means for each level/group.
Choose an FDR cutoff value, default is 10% or 0.10



Select which group is the control to be used for the Fold Change calculation:



The default results from the one-way ANOVA include a fold change column(s), p-value and an FDR column appended on the end of the Final data table. Depending on what else was selected by the user to be printed, other columns may also be appended on the end of the Final data table.

B. One-Way ANOVA with blocking test

This script works very similarly to the above one-way ANOVA script however requires one additional variable column. The additional column should be a descriptor designating which samples belong to which block. A blocking variable can be anything such as replicate number (in paired cases), time, day that sample was processed, patient id, probe array lot etc.

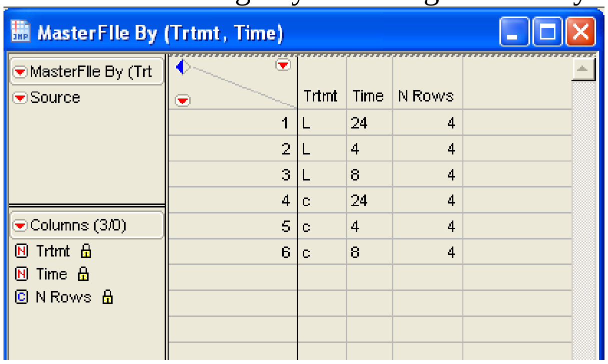
Choose *ANOVA* → *ANOVA1block* on the toolbar menu

C. Two Way ANOVA test

Compute two-way ANOVA: set up 2 categorical columns in master file for 2 factors used in two-way ANOVA.

A. Two-Way ANOVA requires two categorical columns of nominal modeling type.

B. Check balance of design by executing a summary on the two columns:



	Trtmt	Time	N Rows
1	L	24	4
2	L	4	4
3	L	8	4
4	c	24	4
5	c	4	4
6	c	8	4

Design is:

2 factors (Treatment and Time)
4 replicates each factor
Time has 3 levels, Trtmt has 2 levels
6 levels overall

	4hr	**8hr**	**24hr**
c	c4	c8	c24
LPS	LPS4	LPS8	LPS24

This is known as a 2 by 3 factorial design

Short background on a two-way ANOVA:

In experimental designs that incorporate two or more independent variables, the independent variables are called factors, and the designs are called factorial designs. With a two factor design, the analysis yields three pieces of information. There is a test for the **main effect of the first factor (time)**. There is a second test for the **main effect of the second factor (treatment)**. Finally, there is a test that determines if these **two variables interact with one another**. Interactions indicate the joint influence of the two independent variables on the dependent variable. If the variables interact, the effect of one variable depends on the level of the other variable.

(Source: http://espse.ed.psu.edu/statistics/Chapters/Chapter12/Chap12.html)

Two-way ANOVA model:

$\mathbf{Y_{ijk} = E0 + E1_i + E2_j + E12_{ij} + \varepsilon_{ijk}}$

where
i = levels in factor 1
j= levels in factor 2
k= replicate number
ε = the error term
$E1_i$ are the effects due to factor 1
$E2_j$ are the effects due to factor 2

$E12_{ij}$ are the interaction effects

The following constraints are necessary in the two-way ANOVA:

$$\sum_i E1_i = 0$$

$$\sum_j E2_j = 0$$

The two-way ANOVA script prints out E0 (mean), $E1_i$, $E2_j$ and $E12_{ij}$. Due to the constraints, not all the effect parameters are estimated or printed out. Some of the effects will be missing. If you are interested in viewing the missing parameter of your factor of interest, that calculation will be demonstrated by the following example:

Our design has the following set up:

Factor 1: Time – 3 levels

4hr → $E1_i$=4hr

8hr → $E1_i$=8hr

24hr → $E1_i$=24hr

Factor 2: Treatment – 2 levels

LPS → $E2_j$=LPS

C → $E2_j$=c

The script prints out all effects except the first one, thus there is always a missing effect. The missing effect for factor two, Treatment, is $E2_{j=c.}$ Since the two-way ANOVA model has the following constraint $\sum_j E2_j = 0$, the missing effect can be calculated as follows:

$$E2_{j=c} + E2_{j=LPS} = 0$$

$$E2_{j=c} = -E2_{j=LPS}$$

Fold change can be calculated using the effects from the two way ANOVA script. When the factor has 2 levels, the fold change is simply that effect multiplied by 2.

<u>Log Fold change (SFC) = Difference of effects for two levels</u>

In order to calculate fold change, multiply the effect of interest by 2 when the factor has 2 levels:

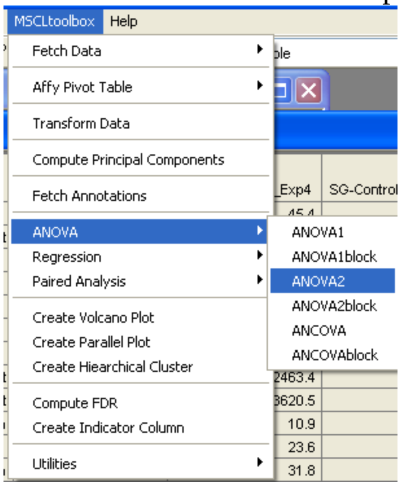
SFC = E2j=LPS - E2j=c

= E2j=LPS – (-E2j=LPS)

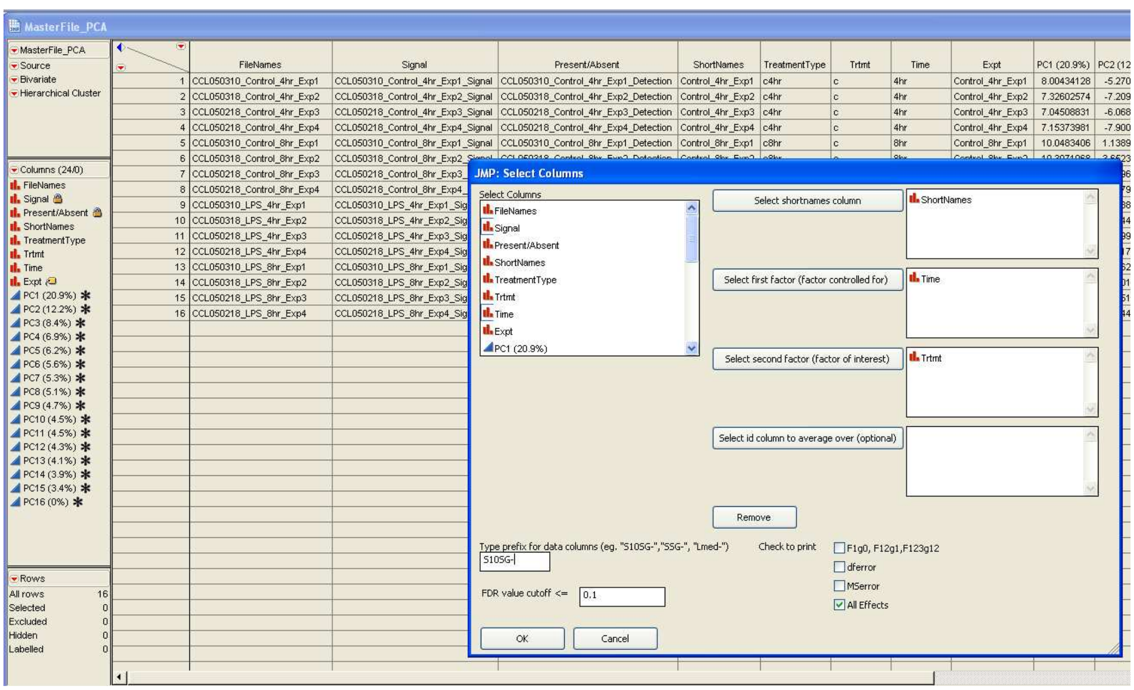
= E2j=LPS + E2j=LPS

= 2* E2j=LPS

Run *ANOVA* → *ANOVA2* to compute a two-way ANOVA from menu toolbar.



Select confounding factor as first factor and treatment of interest as second factor. This order only makes a difference if the design is not balanced. Check print 'All Effects' in order to get the main effects and interaction effect printed from the script. Type in prefix used for transform (ie. S10SG-, Lmed- etc.) and select a cutoff to be used for FDR. Default is 0.10 or 10%.

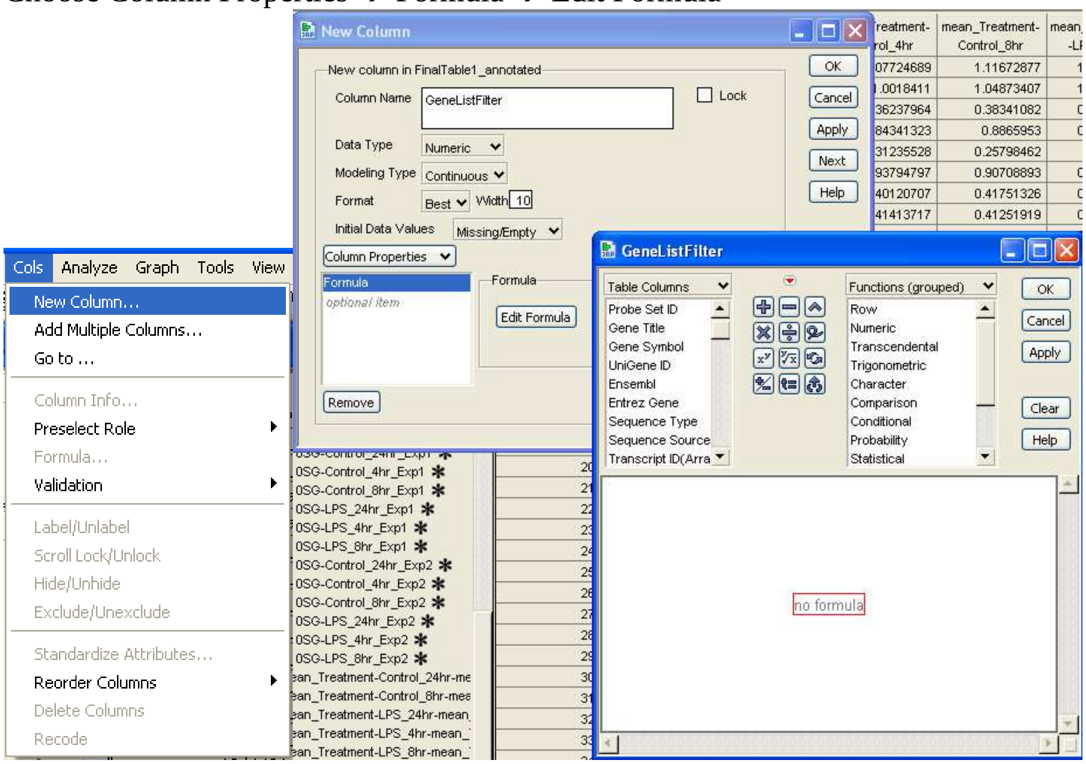


Creating Indicator columns in the Final table

Once you have run your statistical tests, you will then want to choose cutoffs that will filter the data so as to obtain a gene list of interest based on the appropriate filters. Typically these include a False Discovery Rate (FDR) filter, a fold-change filter and a present/absent (PA) call filter. This portion describes how to make these filters using JMP.

Running a statistical test will print out an FDR indicator column, however one may wish to add to that filter, a fold-change cutoff and a PA call cutoff.

Create a new column. Columns → New Column
Type in new Column Name
Choose Column Properties → Formula → Edit Formula

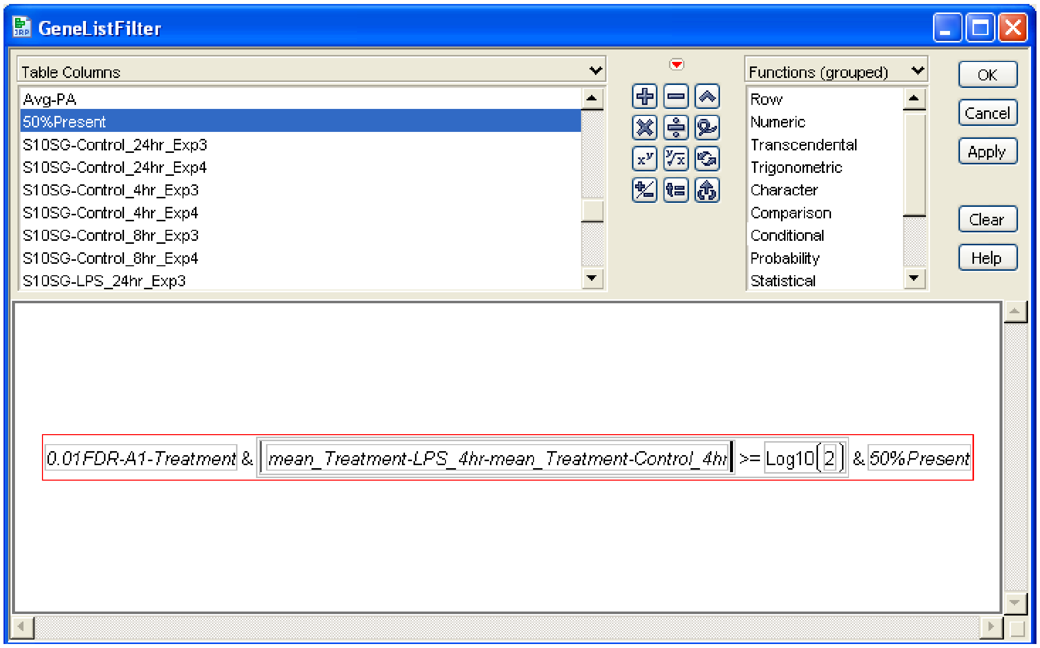


Our new indicator column will contain the following filters:

1% FDR (FDR>=0.01) AND

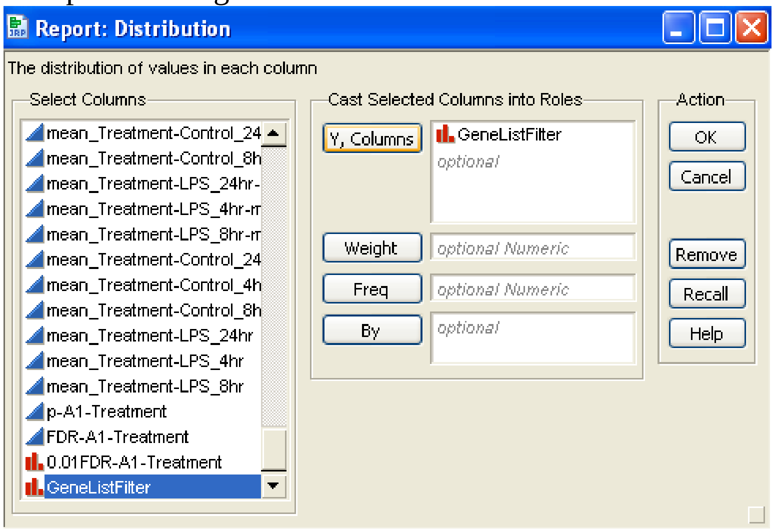
2 fold change or greater (Abs(LFC) >= log10(2)) AND

50% PA (Avg(PA) >= 0.5

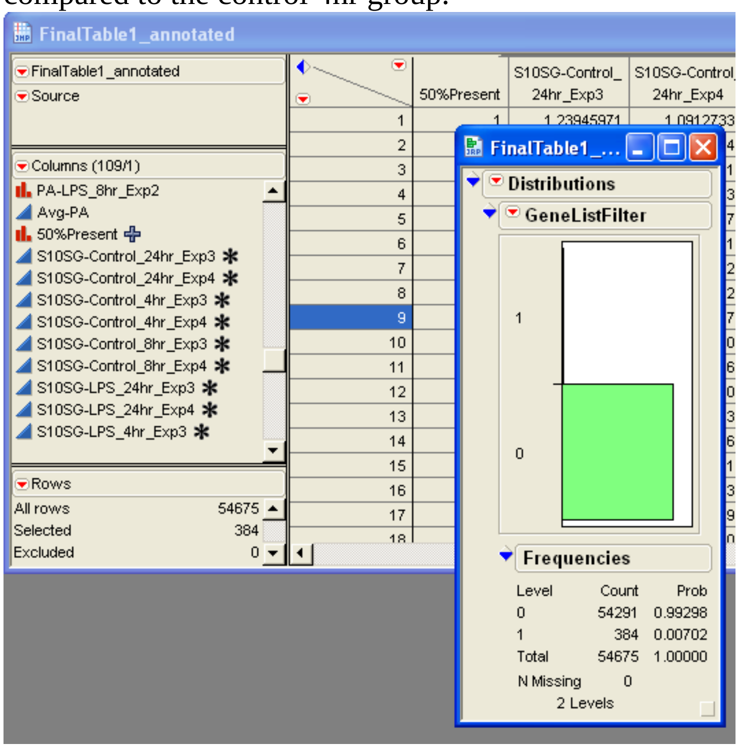


** LFC is computed by the output of the one way ANOVA test. The differences of the means of the LPS4hr group minus the control4hr group is calculated, this is the same as the log fold change of L4hr-c4hr. The absolute value is taken because we want either the 2 fold up genes or the 1/2 fold down genes. Log10(2) is the same as taking the 2 fold change ie. remember we are working in the log10 scale.

Now the fun stuff… What does this give us?
Compute a histogram of the new indicator column filter: *Analyze* → *Distribution*



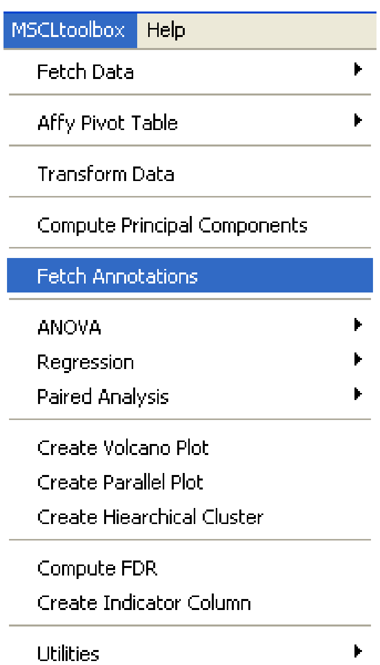
This gives us 384 probe sets that are significantly expressed between the LPS 4hr group compared to the control 4hr group.



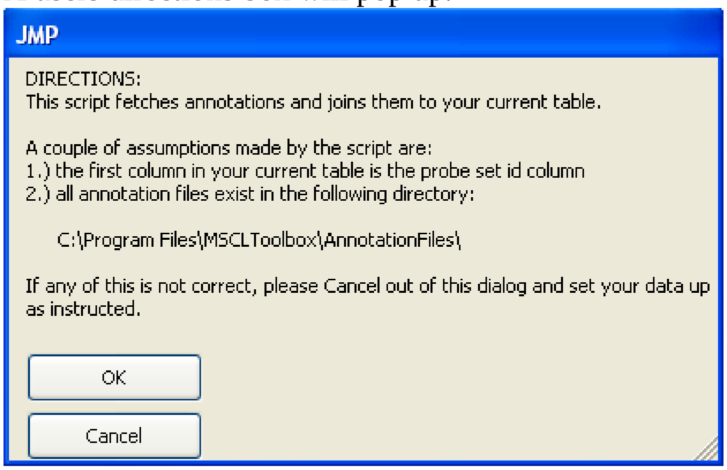
Chapter 8:
Annotating data

The MSCL toolbox provides annotation files for up to 27 Affymetrix gene chips. A new chip can be added easily upon user's request. The annotation files are parsed text files in JMP format that are obtained from the csv files found on the www.netaffx.com website. The annotation files contain annotation information of the following types: gene title/name, gene symbol, Unigene ID, Ensembl ID, Entrez Gene ID, Sequence information, GenBank ID, Chromosomal location information, Swiss Prot ID and Annotation Date. Other information can be obtained upon request.
**If you are running the toolbox locally (i.e. C:\ drive version), then we suggest downloading only the annotation file of interest. It is important to note that the annotation files are very large. You can obtain the current parsed annotation files in JMP format from the following locations:
http://abs.cit.nih.gov/MSCLtoolbox/annotations/AffymetrixAnnotations.html
If you are running the toolbox from your C:\ drive, the you will need to put the annotation file of interest in the following location: C:\MSCLToolbox\AnnotationFiles\
** If you running the toolbox on the msclshare version, then the annotation files are already present and up-to-date.
Once you have the annotation files in the appropriate location, you can proceed below. The *Fetch Annotations* script assumes that the first column in the final table is the probe id column. This is what is used to join in the annotation information.

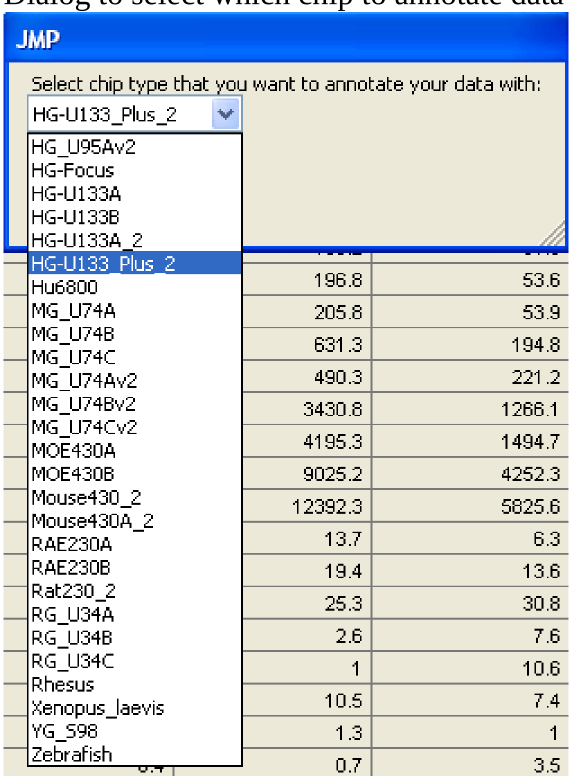
Run *Fetch Annotations* from the toolbar menu



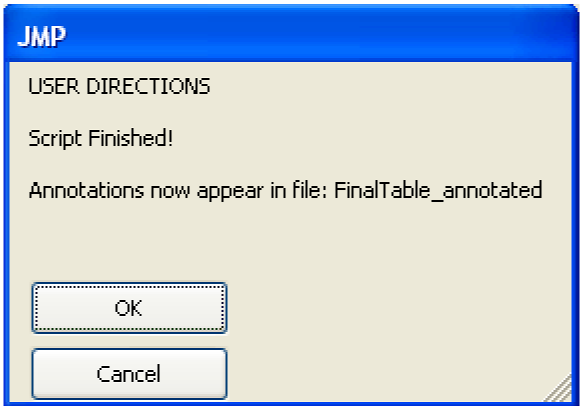
A users directions box will pop up.



Dialog to select which chip to annotate data with pops up. Select appropriate gene chip.



When script is finished running, users directions box will pop up. Annotations are now at the beginning of the Final file. Please save new file in appropriate location.



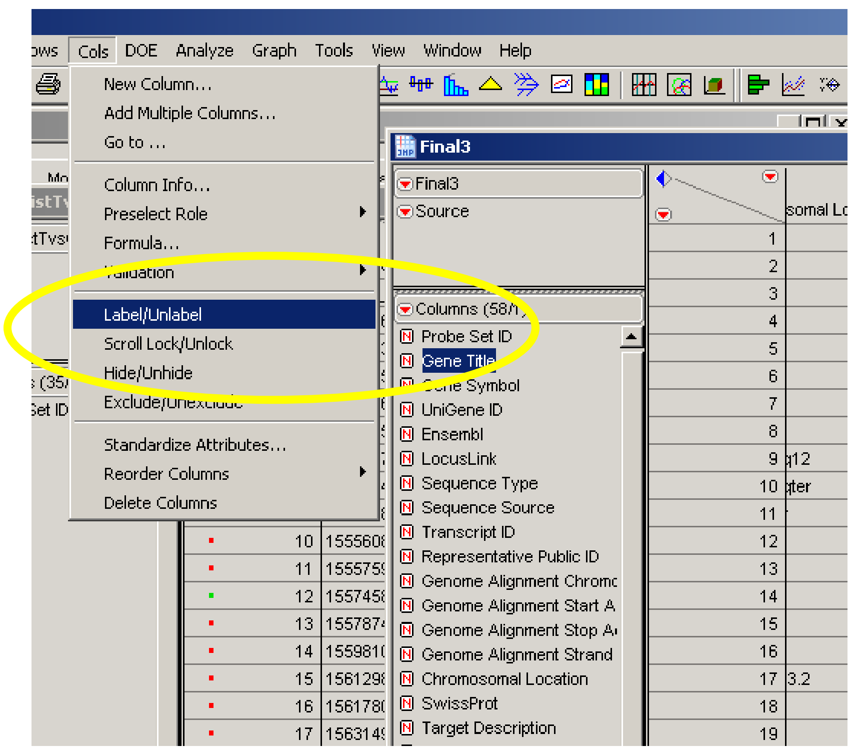
The new table is a copy of your Final table with annotations attached. You will want to save this new table.

Chapter 9:
Data exploration and visualization tools

A. Label
B. Volcano Plot
C. Selecting genes
D. Making a Subset
E. Hierarchical Cluster
F. Parallel Plots
G. Summary with Venn Diagram

A. Turn Column Labels on

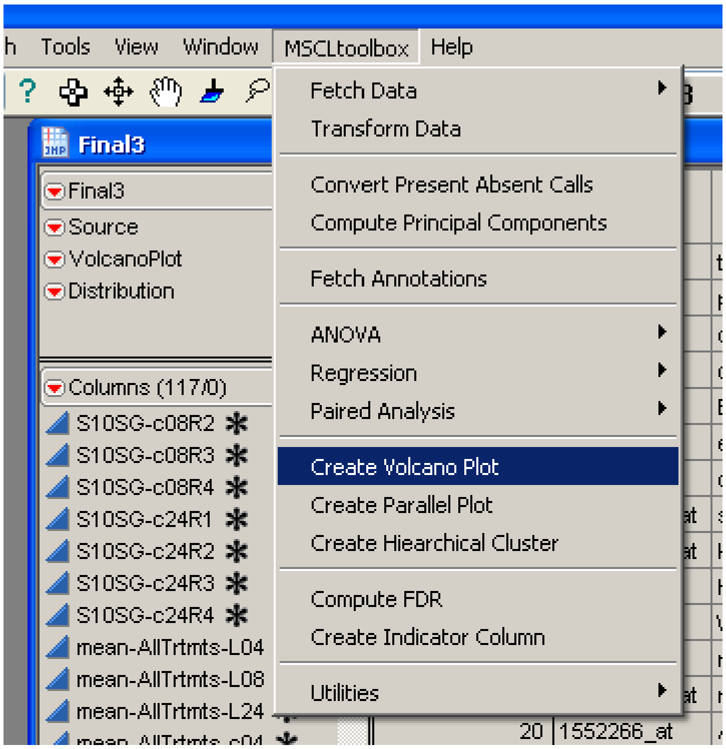
This allows one to see the label when hovering over a record in any visualization tool.



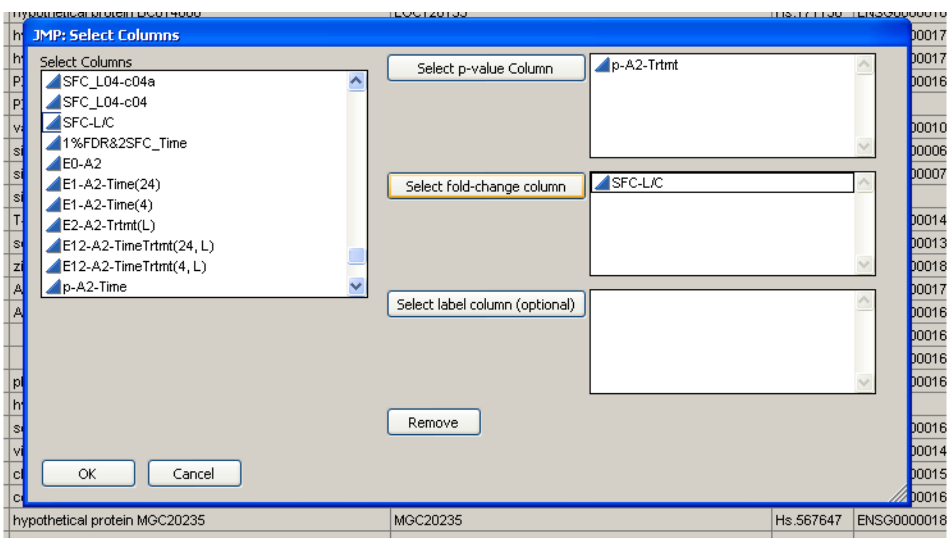
Select column to turn on label then choose **Cols→Label→Unlabel**

B. Make a Volcano Plot – Log P-value by Log Fold Change

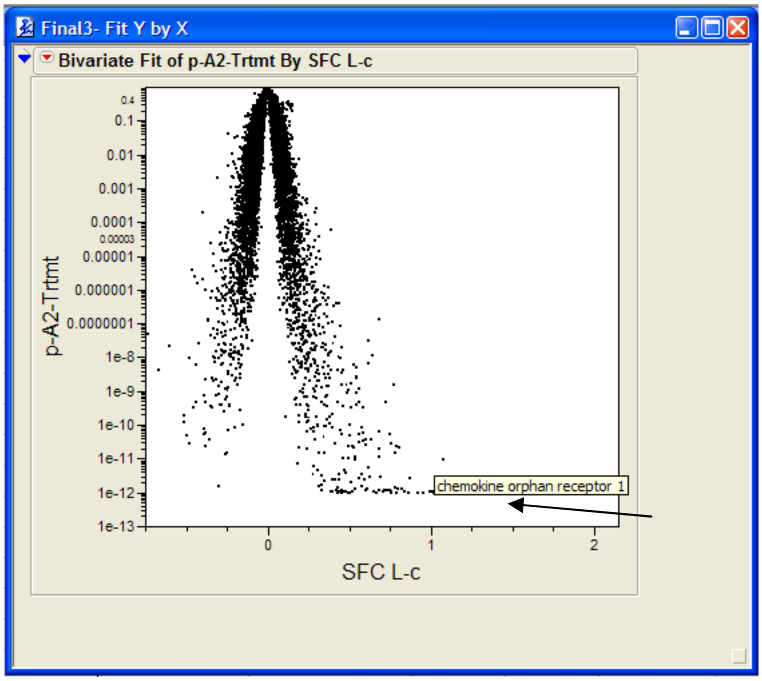
Select **MSCLtoolbox → Create Volcano Plot**



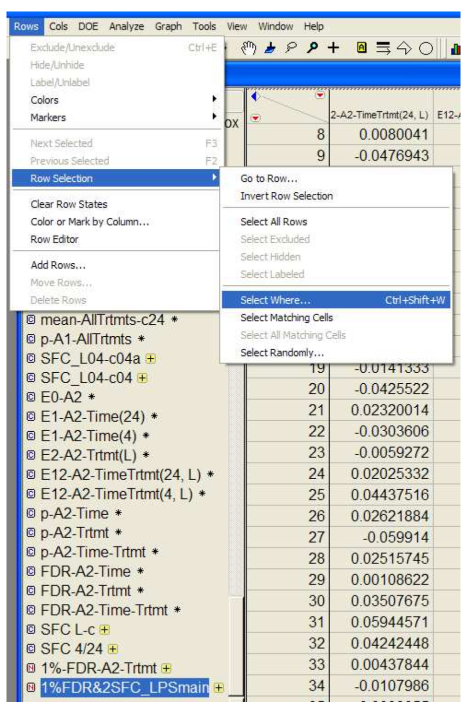
You will be prompted to choose the P-value and Fold change columns.

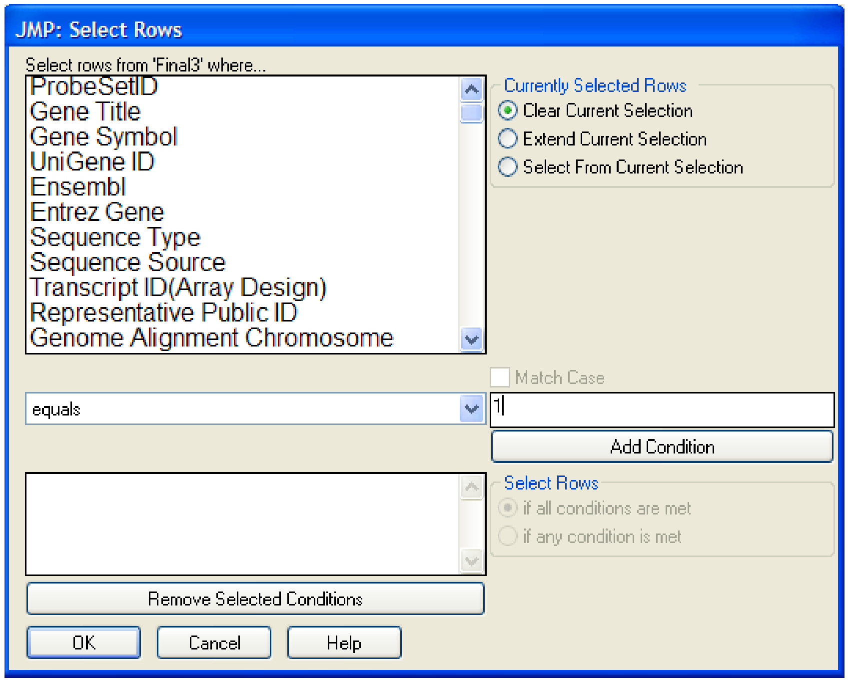


Circle significant genes with a low pvalue and high fold change. Color up genes red and down genes green using Row properties tool.

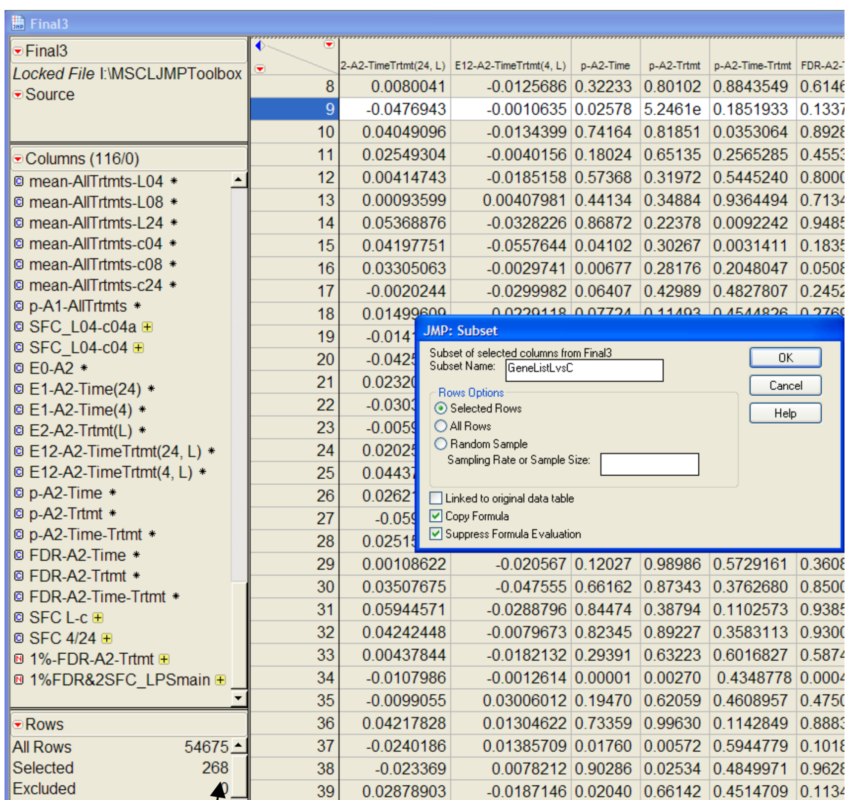


C. Select genes that pass filter by selecting where indicator column equals 1.



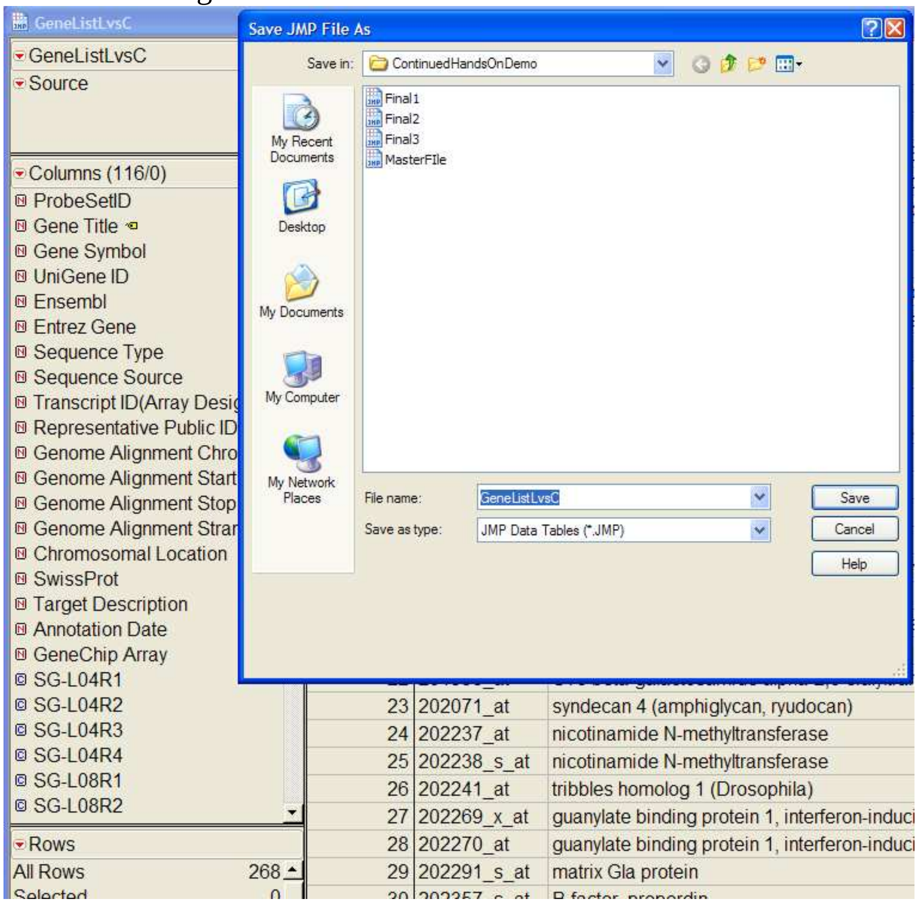


D. Make Subset of gene list and save

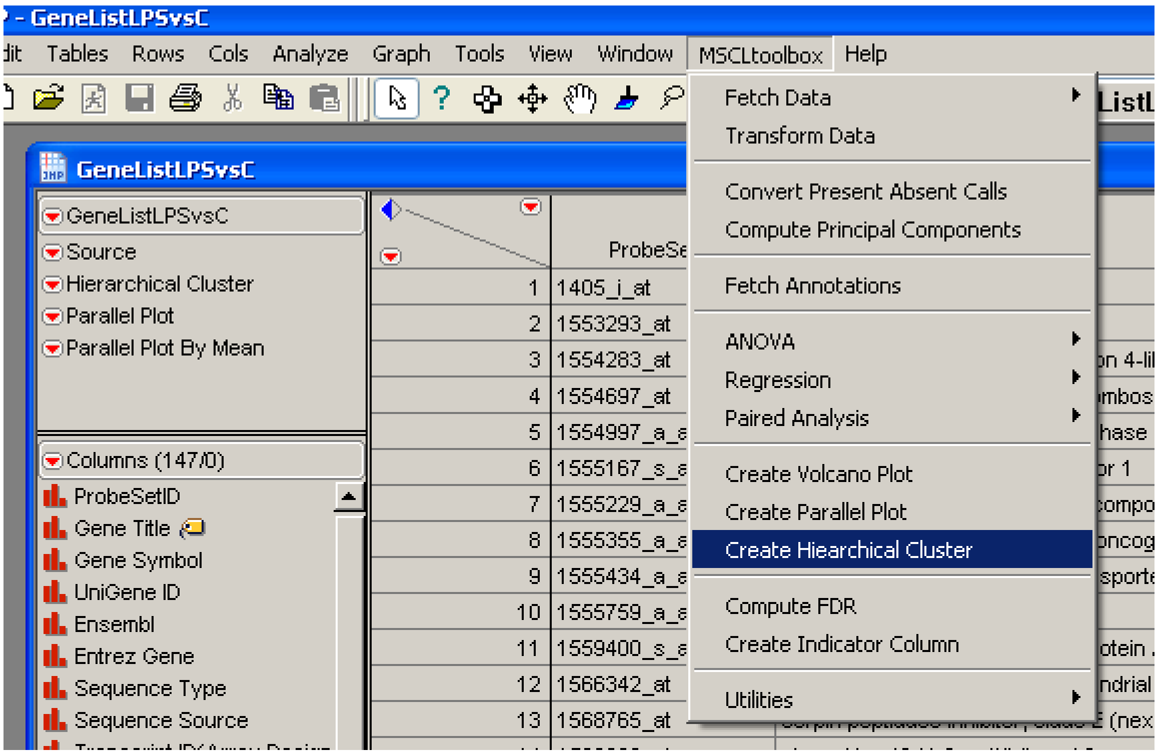


Make sure genes are selected and then make subset, click OK.

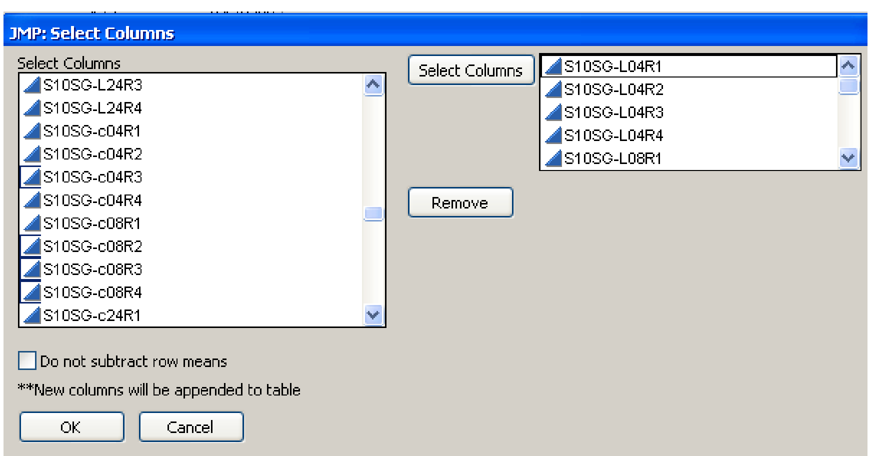
Save subset of genes as “GeneListLvsC”.



E. Make Hierarchical Cluster (HC) – by selecting **MSCLtoolbox→Create Hierarchical Cluster** on the subset of selected genes. **Note:** do not try to cluster the entire chip (20,000+ probe sets). The JMP program cannot handle that many probe sets at once.

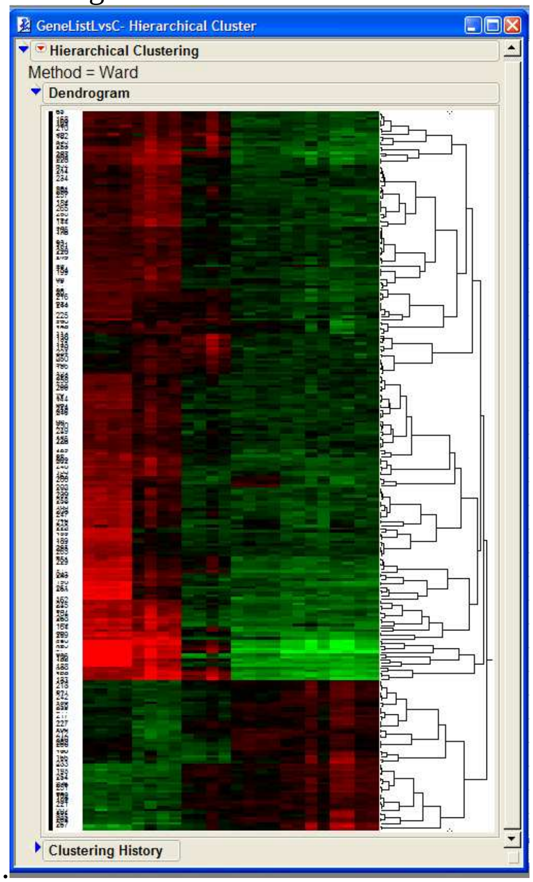


Choose the transform signal columns



Then click OK

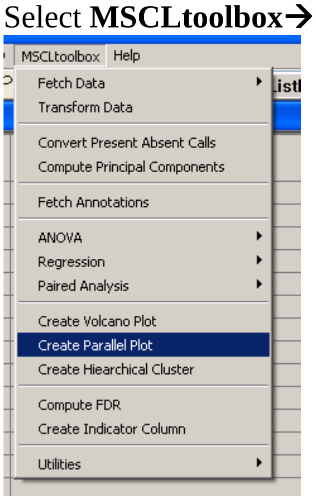
Investigate Hierarchical cluster and look for clusters of interest.



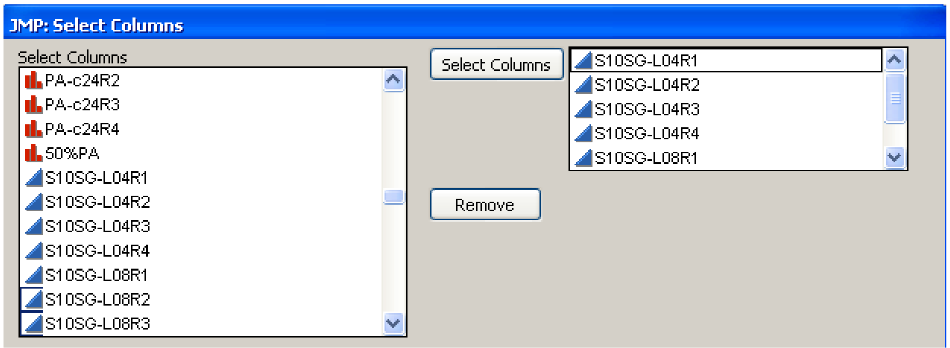
.

F. Make Parallel Gene plots

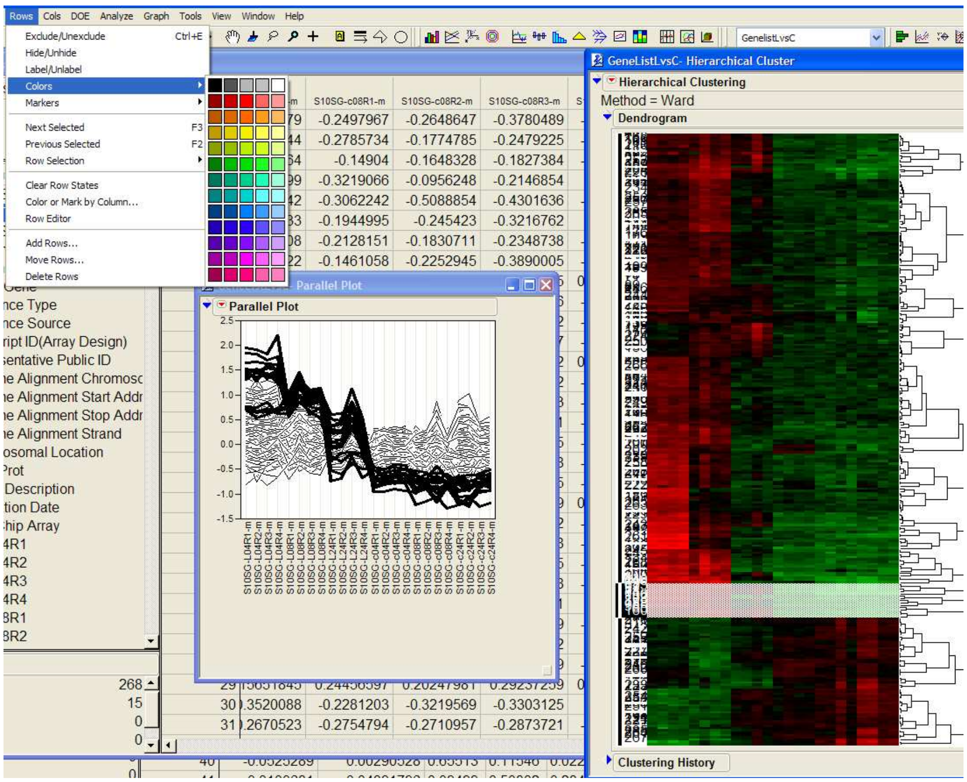
Select **MSCLtoolbox→ Create Parallel Plot**



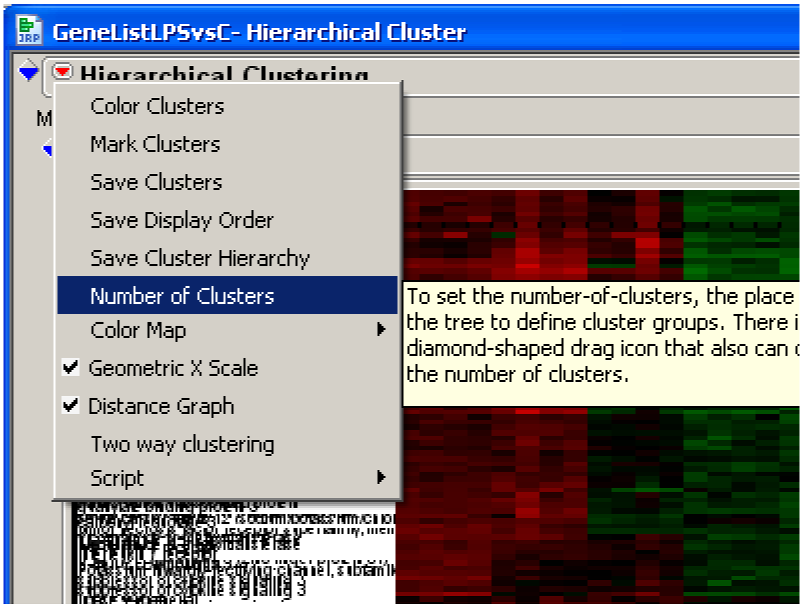
Then select the normalized columns to be plotted on parallel plot



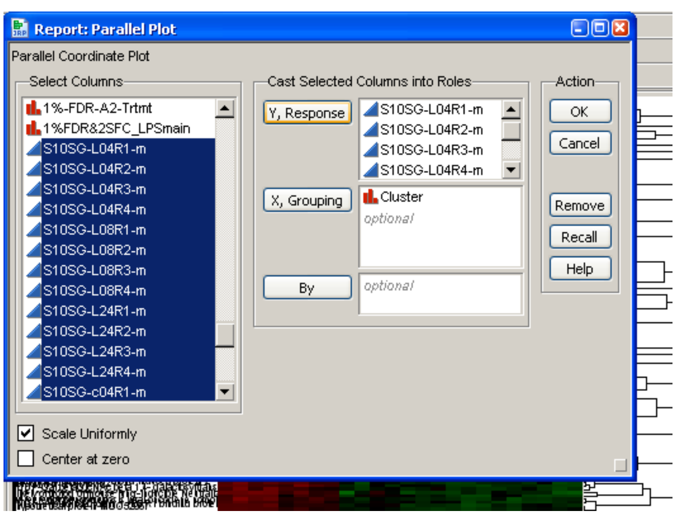
Output can be viewed next to Hierarchical cluster picture and can be interactively interrogated. It is helpful to color genes of interest.



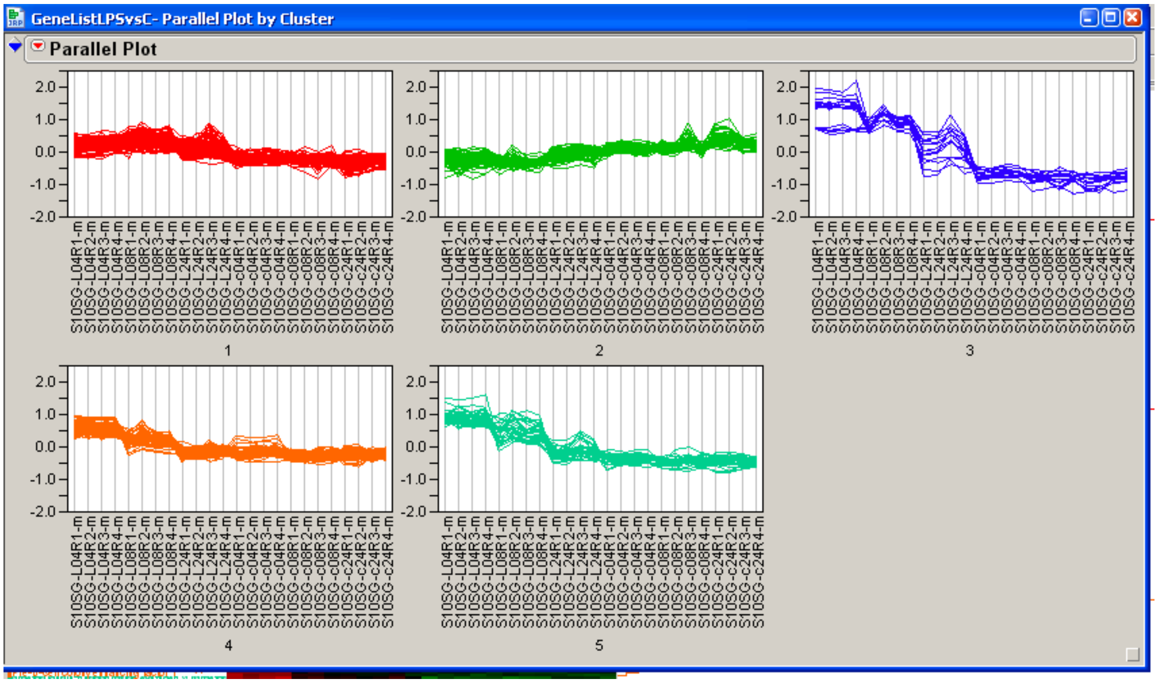
We can now view the Parallel Plot with the data colored by cluster. To do this, first set the number of clusters to 5 by clicking on the red triangle → **Number of Clusters** and then enter 5.



Now save the cluster number in the data table by clicking the red triangle → **Save Clusters.** This will add a column to the data table.
At this point color the data by cluster by going to **Rows→Color By Column** and select the cluster column. For the final output select **Graph→ Parallel Plot.** Select the normalized signal columns with the mean subtracted off for the Y Response, Cluster for the X, Grouping and check the Scale Uniformly box.



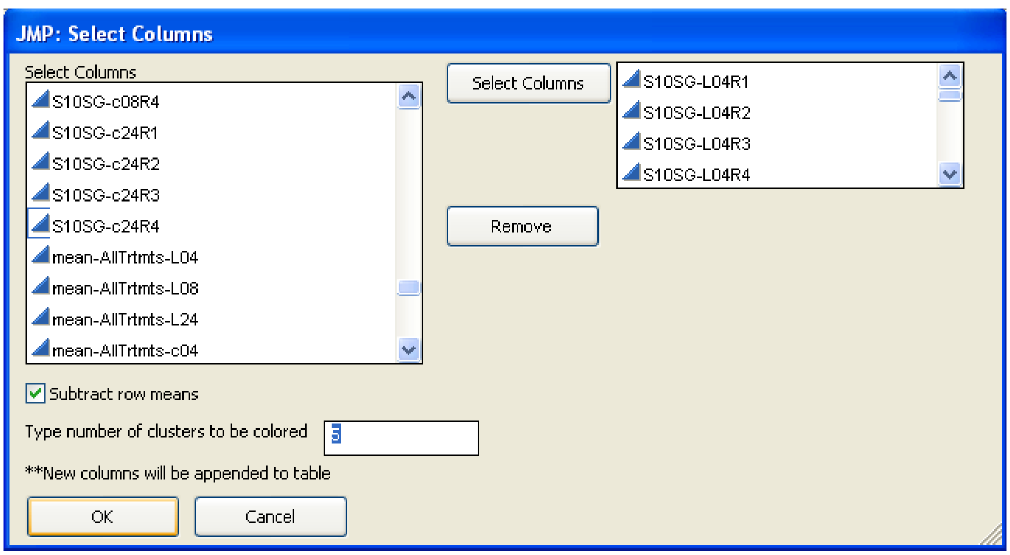
The result will be similar to the graphs below.



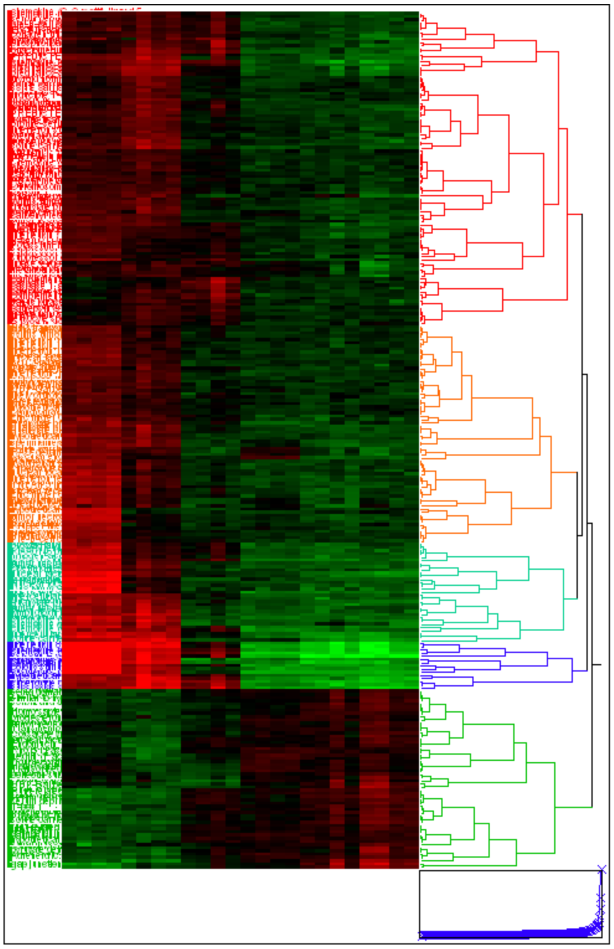
G. Make Hierarchical cluster and parallel plot together with one script

Select **MSCLtoolbox→ Create Hierarchical Cluster and Parallel Plot**

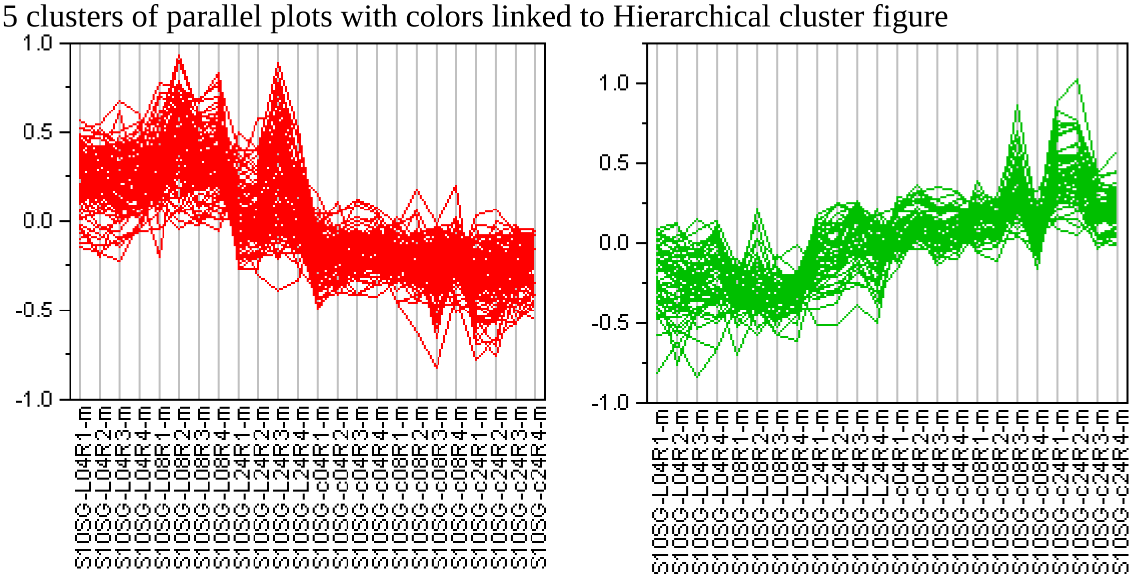
Select columns to be displayed in plots. Script will subtract means if check box is checked. Choose the number of clusters to save and display in the parallel plots.

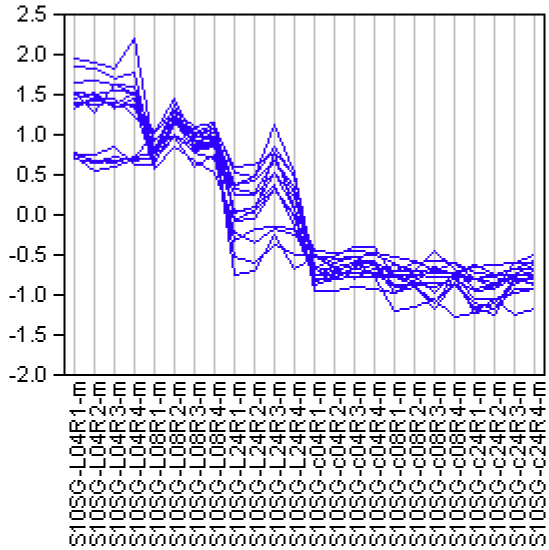


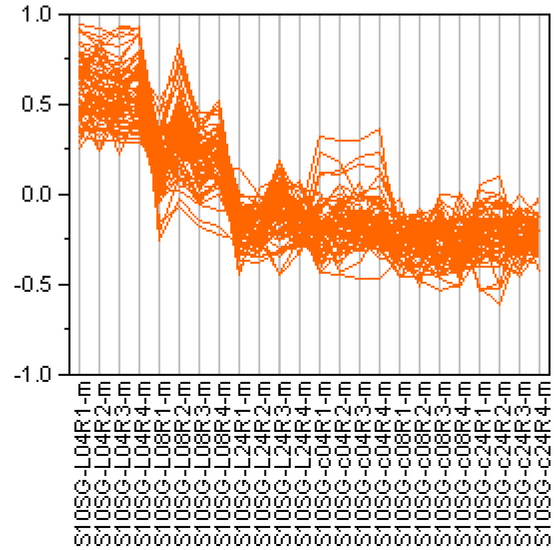
Output:
Hierarchical cluster figure with 5 clusters colored

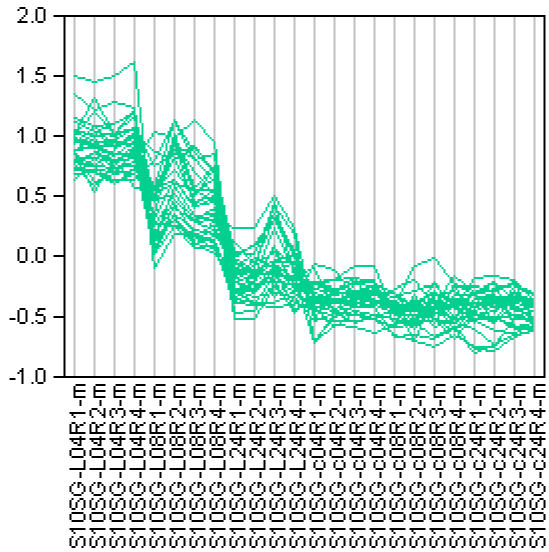


5 clusters of parallel plots with colors linked to Hierarchical cluster figure



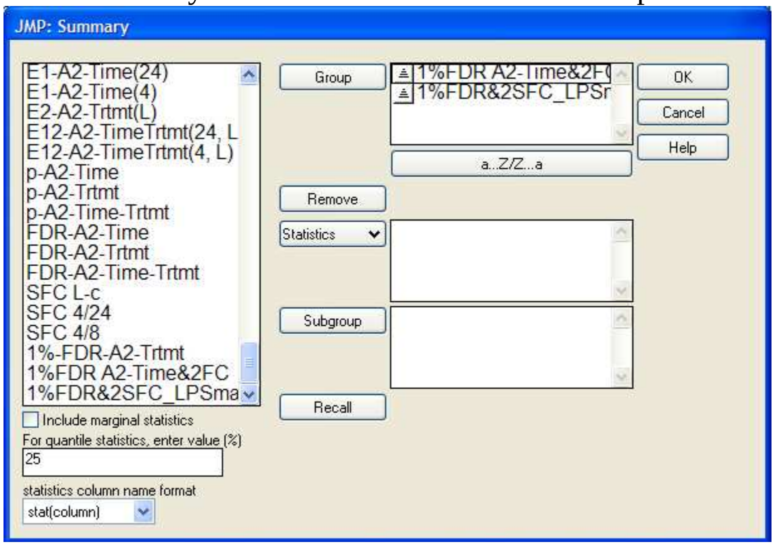

2.5
2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.5
-2.0
S10SG-L04R1-m
S10SG-L04R2-m
S10SG-L04R3-m
S10SG-L04R4-m
S10SG-L08R1-m
S10SG-L08R2-m
S10SG-L08R3-m
S10SG-L08R4-m
S10SG-L24R1-m
S10SG-L24R2-m
S10SG-L24R3-m
S10SG-L24R4-m
S10SG-c04R1-m
S10SG-c04R2-m
S10SG-c04R3-m
S10SG-c04R4-m
S10SG-c08R1-m
S10SG-c08R2-m
S10SG-c08R3-m
S10SG-c08R4-m
S10SG-c24R1-m
S10SG-c24R2-m
S10SG-c24R3-m
S10SG-c24R4-m

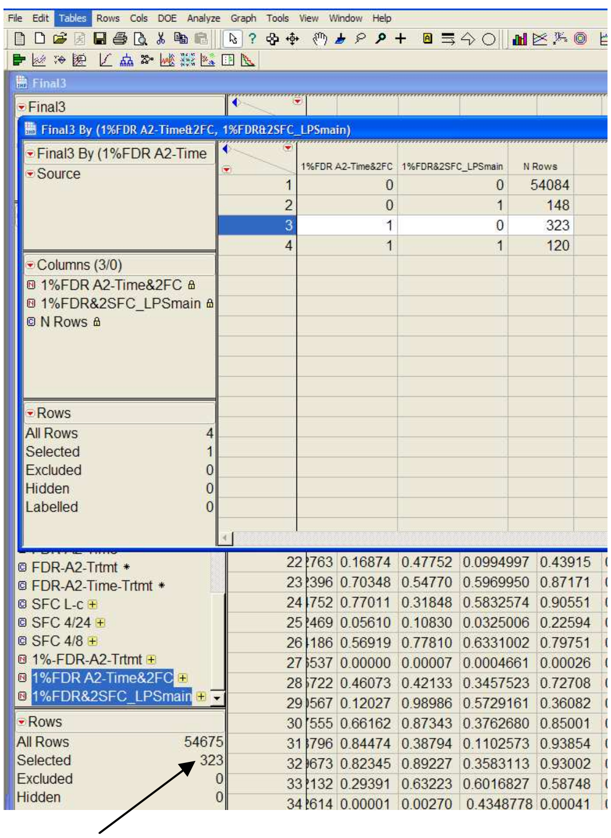

1.0
0.5
0.0
-0.5
-1.0
S10SG-L04R1-m
S10SG-L04R2-m
S10SG-L04R3-m
S10SG-L04R4-m
S10SG-L08R1-m
S10SG-L08R2-m
S10SG-L08R3-m
S10SG-L08R4-m
S10SG-L24R1-m
S10SG-L24R2-m
S10SG-L24R3-m
S10SG-L24R4-m
S10SG-c04R1-m
S10SG-c04R2-m
S10SG-c04R3-m
S10SG-c04R4-m
S10SG-c08R1-m
S10SG-c08R2-m
S10SG-c08R3-m
S10SG-c08R4-m
S10SG-c24R1-m
S10SG-c24R2-m
S10SG-c24R3-m
S10SG-c24R4-m


2.0
1.5
1.0
0.5
0.0
-0.5
-1.0
S10SG-L04R1-m
S10SG-L04R2-m
S10SG-L04R3-m
S10SG-L04R4-m
S10SG-L08R1-m
S10SG-L08R2-m
S10SG-L08R3-m
S10SG-L08R4-m
S10SG-L24R1-m
S10SG-L24R2-m
S10SG-L24R3-m
S10SG-L24R4-m
S10SG-c04R1-m
S10SG-c04R2-m
S10SG-c04R3-m
S10SG-c04R4-m
S10SG-c08R1-m
S10SG-c08R2-m
S10SG-c08R3-m
S10SG-c08R4-m
S10SG-c24R1-m
S10SG-c24R2-m
S10SG-c24R3-m
S10SG-c24R4-m

H. Summary of data

Make summary of nominal columns to see overlap.

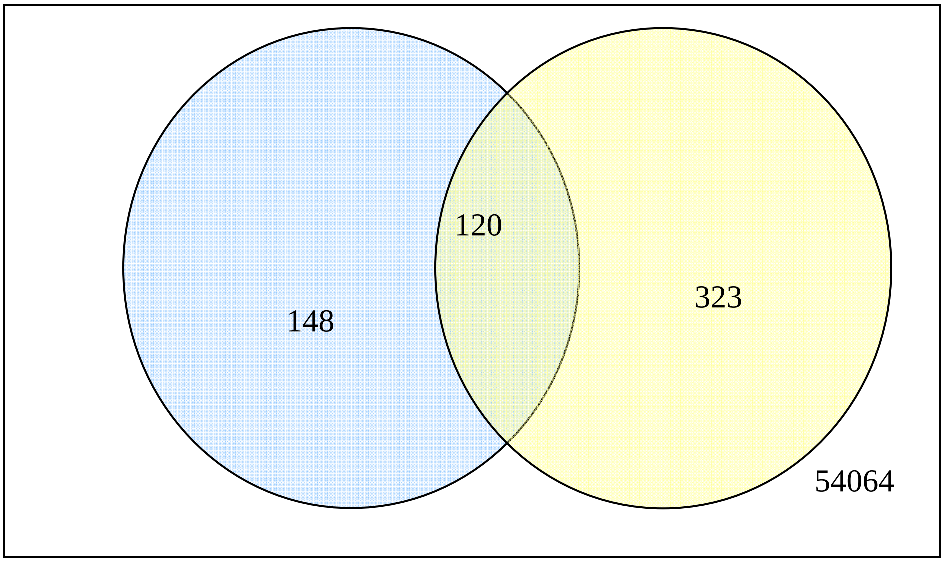




Select genes that are selected by one genelist and not on another then make a subset.

A summary of this nature can be used to easily create a Venn diagram.

Using the numbers from the summary table the following diagram was created.


120
148
323
54064

Chapter 4:

Analysis of Affymetrix Exon chips

ExonANOVAnested script:

The ANOVA nested model used for Exon Array analysis is:

$$y_{ijk} = \mu + A_i + \beta_{j(i)} + C_k + AC_{ik} + \varepsilon_{ijk}$$

<u>2 fixed, one random effect</u>

A_i	Treatment effect (fixed)
$\beta_{j(i)}$	Sample within treatment effect (random)
C_k	Exon effect (fixed)
AC_{ik}	Treatment-exon interaction effect (fixed)
ε_{ijk}	error term

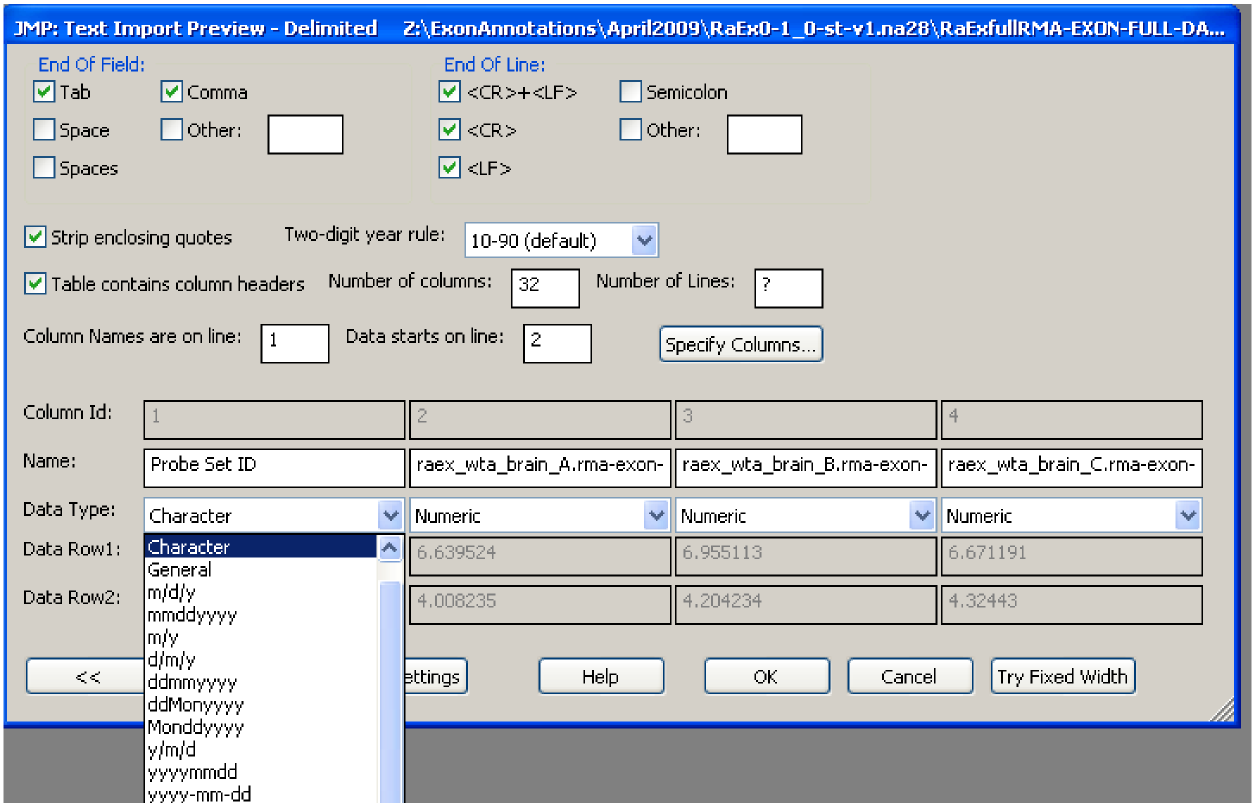
The ANOVA Table

Source	SS	df	E(MS)	F-Test
A, fixed TREATMENT	$bc\sum_{i=1}^{a}(y_{i..} - y_{...})^2$	$a-1$	$\frac{bc\sum_{i=1}^{a} A_i^2}{(a-1)} + c\sigma_B^2 + \sigma^2$	$\frac{MS_A}{MS_{B(A)}}$
B, random SAMPLE	$c\sum_{i=1}^{a}\sum_{j(i)=1}^{b}(y_{ij(i).} - y_{i..})^2$	$a(b-1)$	$c\sigma_B^2 + \sigma^2$	$\frac{MS_{B(A)}}{MS_{error}}$
C, fixed EXON	$ab\sum_{k=1}^{c}(y_{..k} - y_{...})^2$	$c-1$	$\frac{ab\sum_{k=1}^{c} C_k^2}{(c-1)} + \sigma^2$	$\frac{MS_C}{MS_{error}}$
AC, fixed TRT x EXON	$b\sum_{i=1}^{a}\sum_{k=1}^{c}(y_{i.k} - y_{i..} - y_{..k} + y_{...})^2$	$(a-1)(c-1)$	$\frac{b\sum_{i=1}^{a}\sum_{k=1}^{c}(AC)_{ik}^2}{(a-1)(c-1)} + \sigma^2$	$\frac{MS_{AC}}{MS_{error}}$
error	$\sum_{i=1}^{a}\sum_{j(i)=1}^{b}\sum_{k=1}^{c}(y_{ij(i)k} - y_{i.k} - y_{ij(i).} + y_{i..})^2$	$a(b-1)(c-1)$	σ^2	
corrected Total	$\sum_{i=1}^{a}\sum_{j(i)=1}^{b}\sum_{k=1}^{c}(y_{ij(i)k} - y_{...})^2$	$abc-1$		

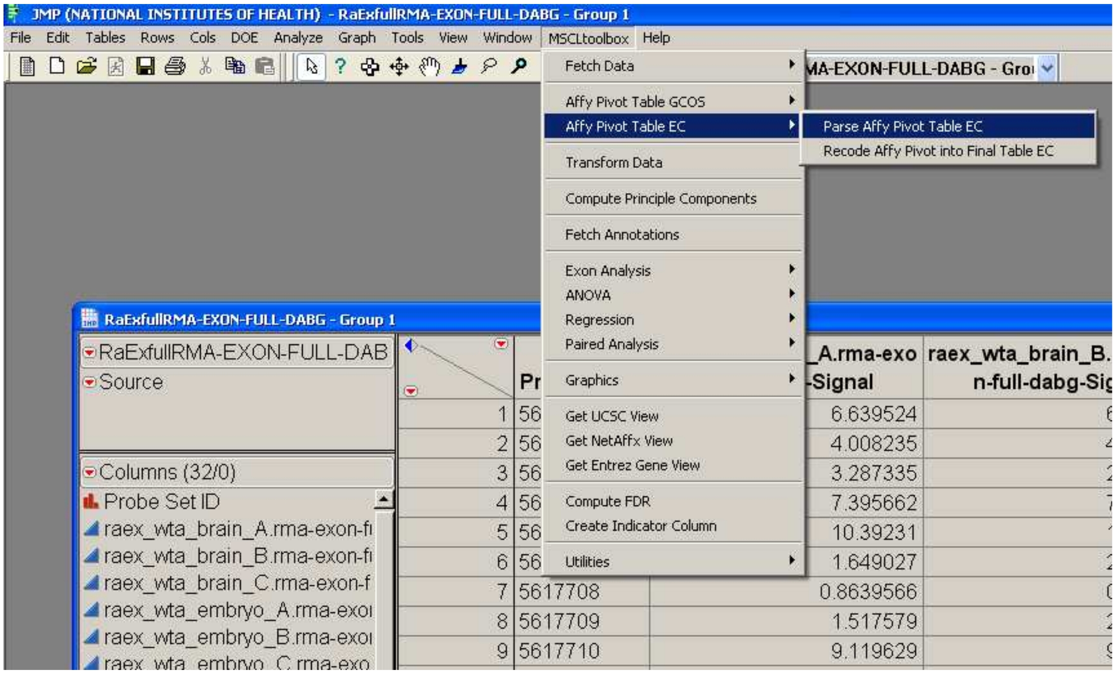
The A factor can be looked at to determine differential gene expression.
The AC factor can be looked at to determine alternative splicing.

Steps for analyzing data from the three Affymetrix Exon arrays:
Human, Mouse, Rat

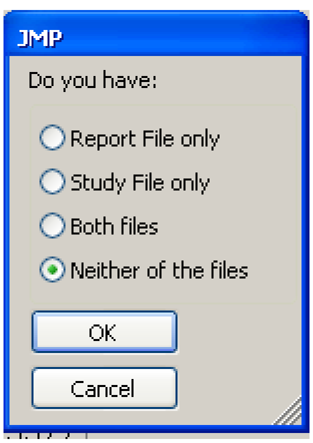
1. Import processed data from EC into JMP using the Text Import Preview Option

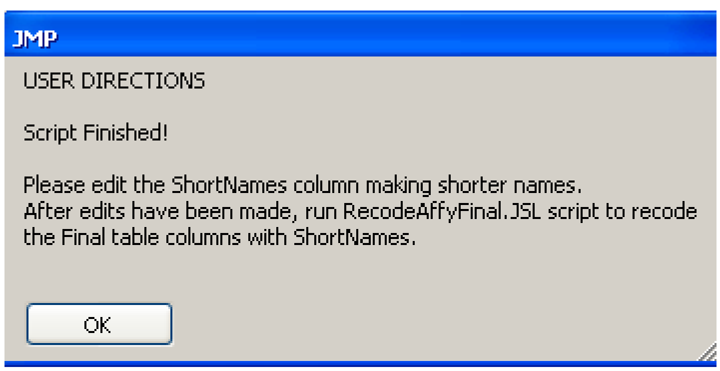


2. Parse Affy Pivot Table from EC into Final Table and create the Master File that links up the Signal columns to rows in the Master File



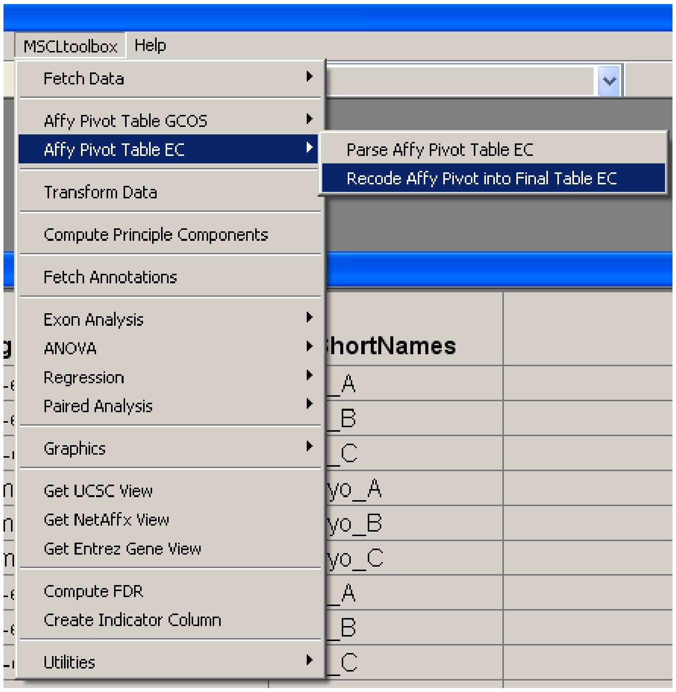
Dialog box to the user :
Did user export a Report, Study, Both or Neither files from EC? When script is finished, a USER DIRECTIONS box will be displayed.



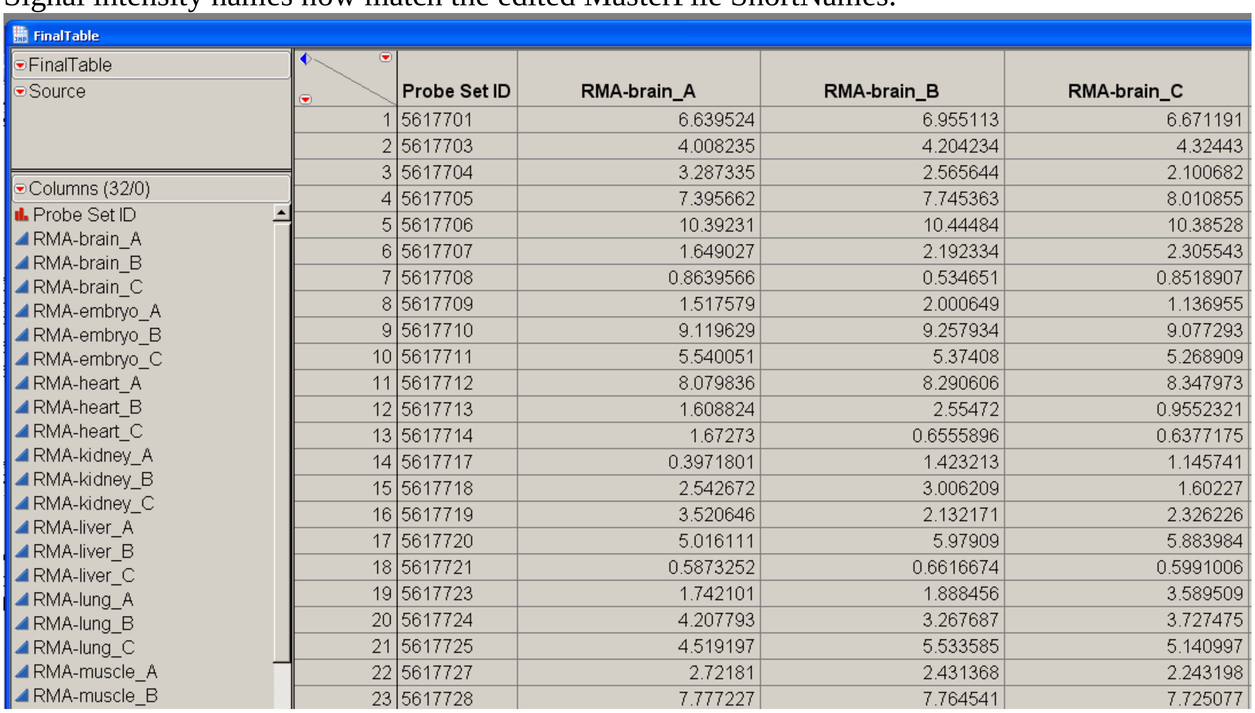


Edit ShortNames in MasterFile creating names that are more succinct to be added back to the Final table.

3. Recode the Affy Pivot table into FinalTable extracting the newly created ShortNames entries to be added as column names in the FinalTable.

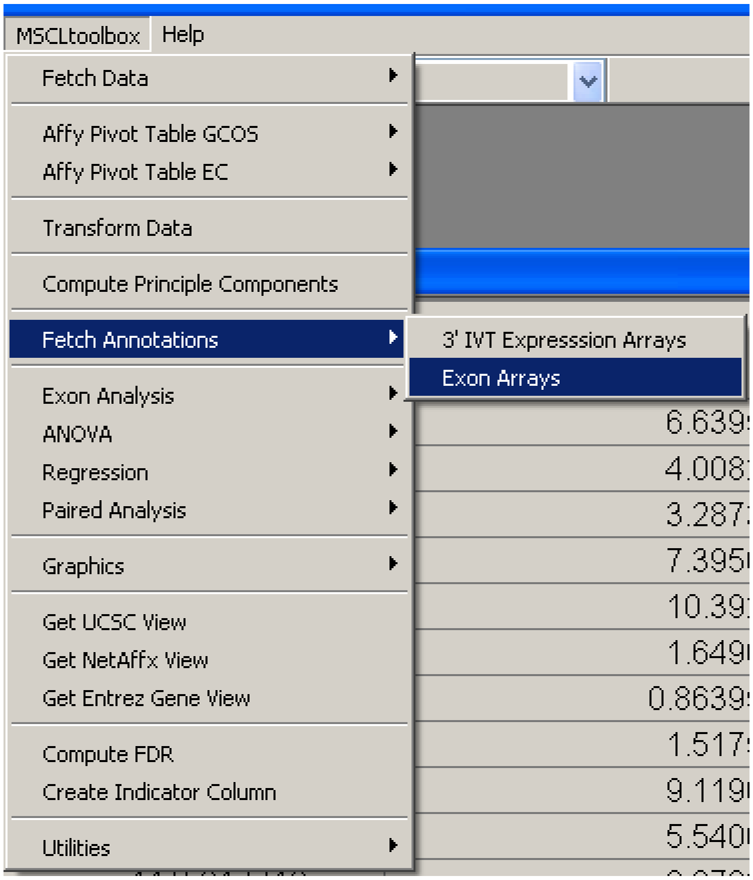


Signal intensity names now match the edited MasterFile ShortNames.

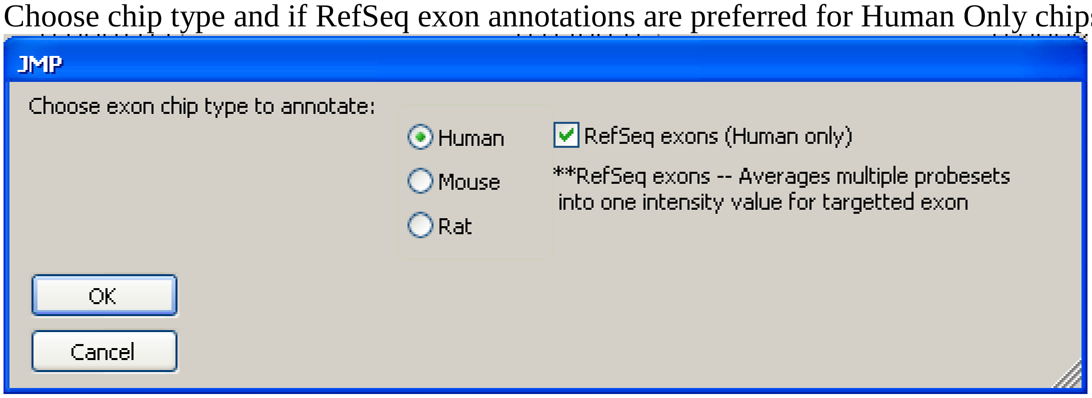


	Probe Set ID	RMA-brain_A	RMA-brain_B	RMA-brain_C
1	5617701	6.639524	6.955113	6.671191
2	5617703	4.008235	4.204234	4.32443
3	5617704	3.287335	2.565644	2.100682
4	5617705	7.395662	7.745363	8.010855
5	5617706	10.39231	10.44484	10.38528
6	5617707	1.649027	2.192334	2.305543
7	5617708	0.8639566	0.534651	0.8518907
8	5617709	1.517579	2.000649	1.136955
9	5617710	9.119629	9.257934	9.077293
10	5617711	5.540051	5.37408	5.268909
11	5617712	8.079836	8.290606	8.347973
12	5617713	1.608824	2.55472	0.9552321
13	5617714	1.67273	0.6555896	0.6377175
14	5617717	0.3971801	1.423213	1.145741
15	5617718	2.542672	3.006209	1.60227
16	5617719	3.520646	2.132171	2.326226
17	5617720	5.016111	5.97909	5.883984
18	5617721	0.5873252	0.6616674	0.5991006
19	5617723	1.742101	1.888456	3.589509
20	5617724	4.207793	3.267687	3.727475
21	5617725	4.519197	5.533585	5.140997
22	5617727	2.72181	2.431368	2.243198
23	5617728	7.777227	7.764541	7.725077

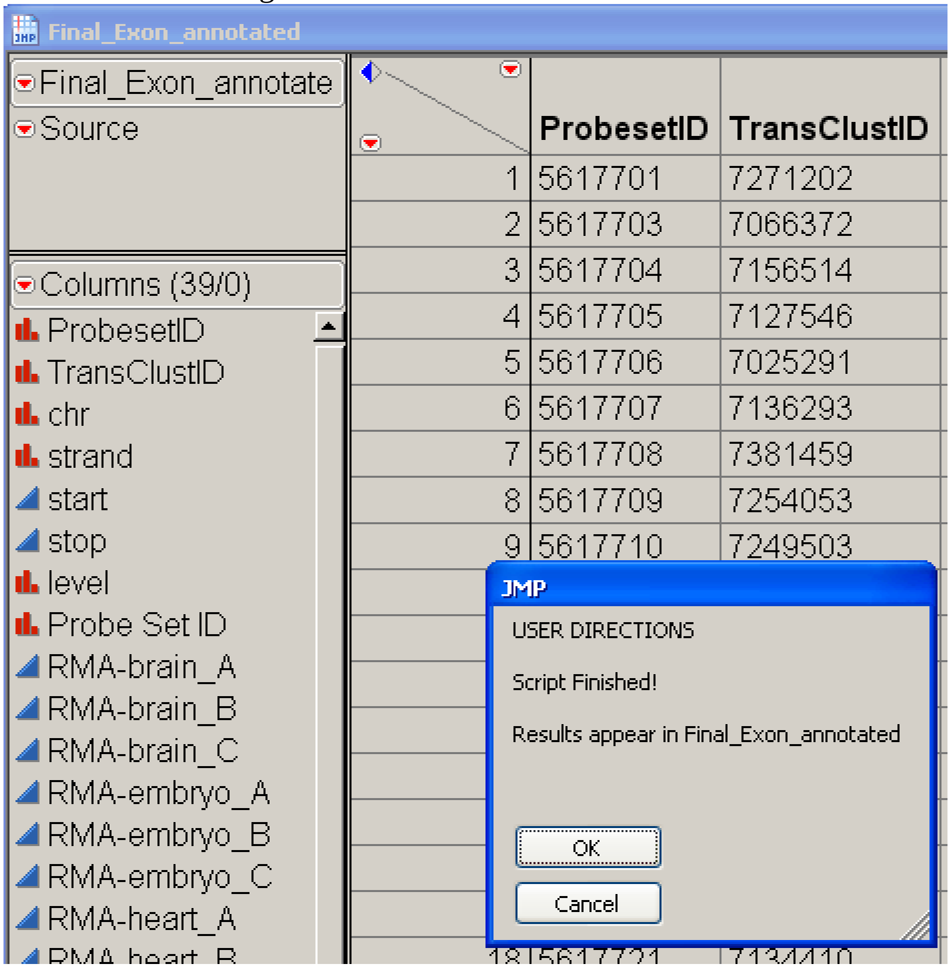
4. Annotate Exon Array to include max Intensity and range over all tissues. Other annotations include, GeneID, chromosome number, strand and start and stop location for each exon.



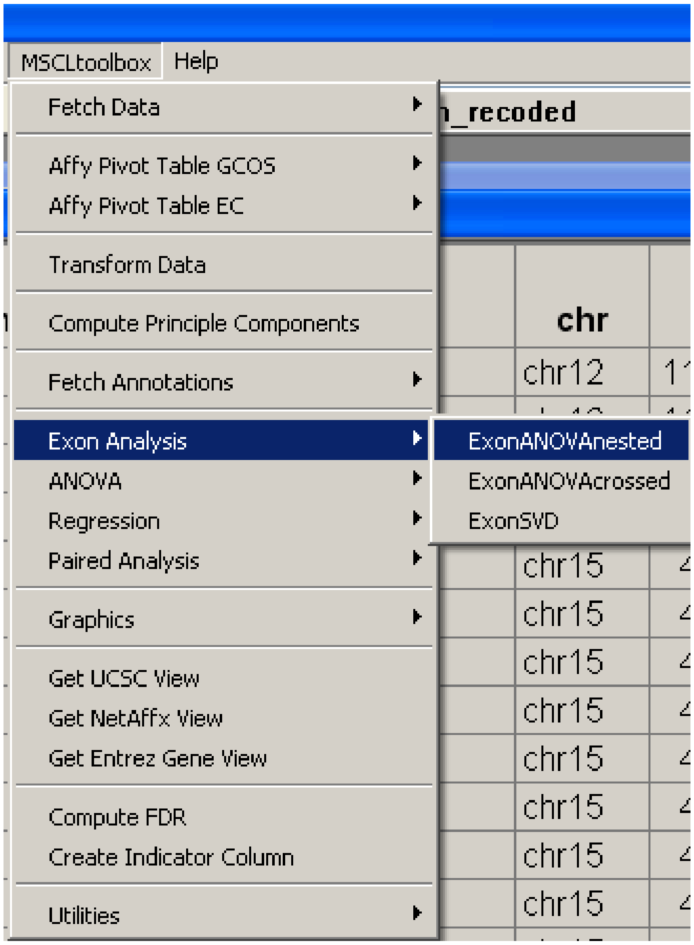
Choose chip type and if RefSeq exon annotations are preferred for Human Only chips.



New annotations now appear in the Final table. *** User must have access to the mscltoolbox.cit.nih.gov/MSCLtoolbox location available to all NIH users.

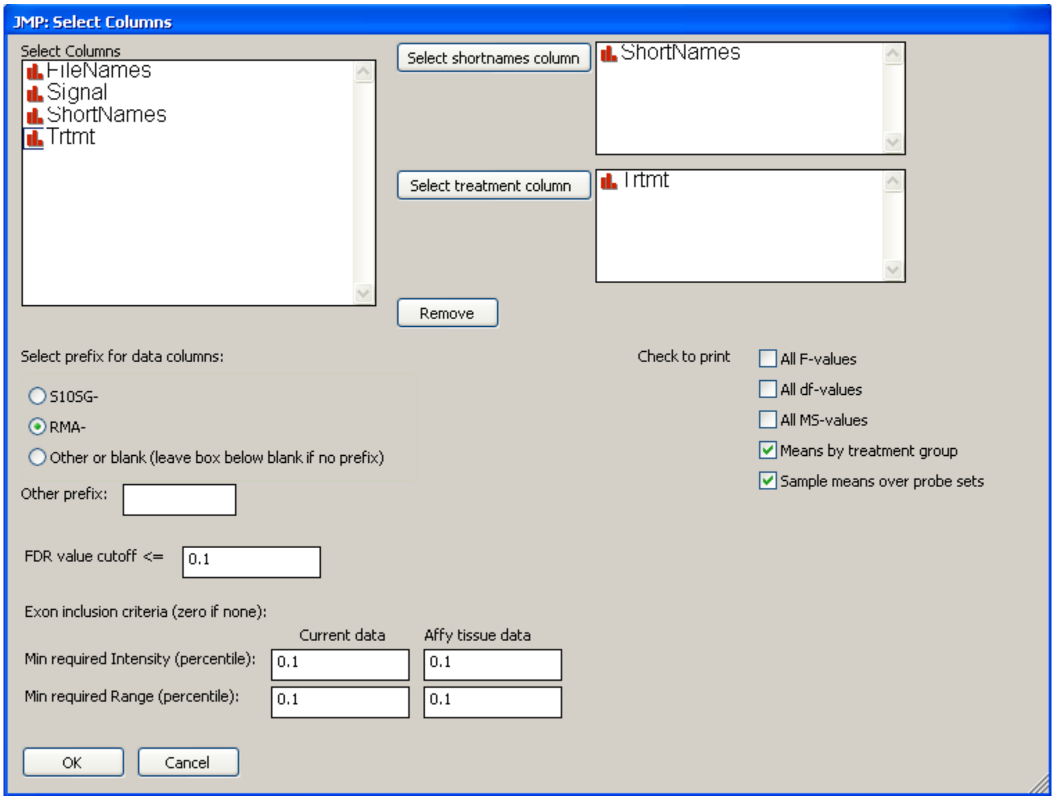


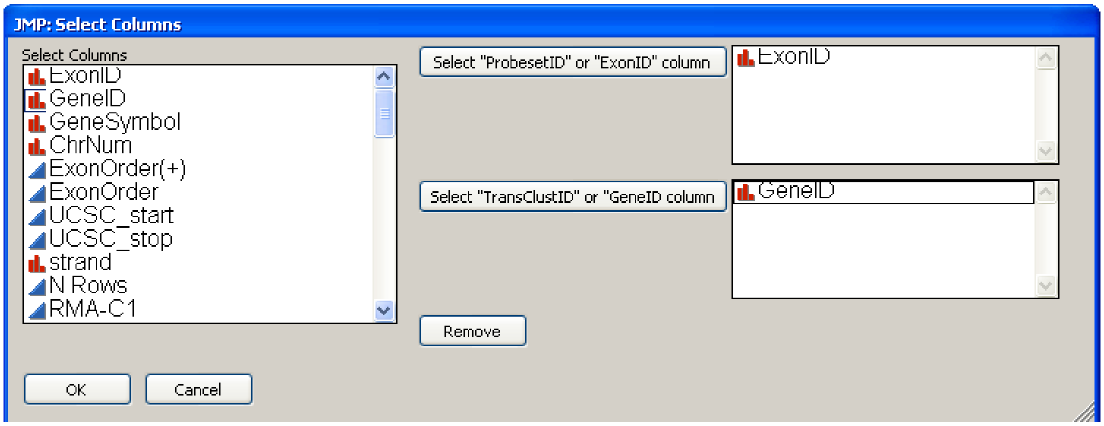
5. Run ExonANOVAnested. Result of running this script are two files: Final_ExonLevel and Final_GeneLevel.



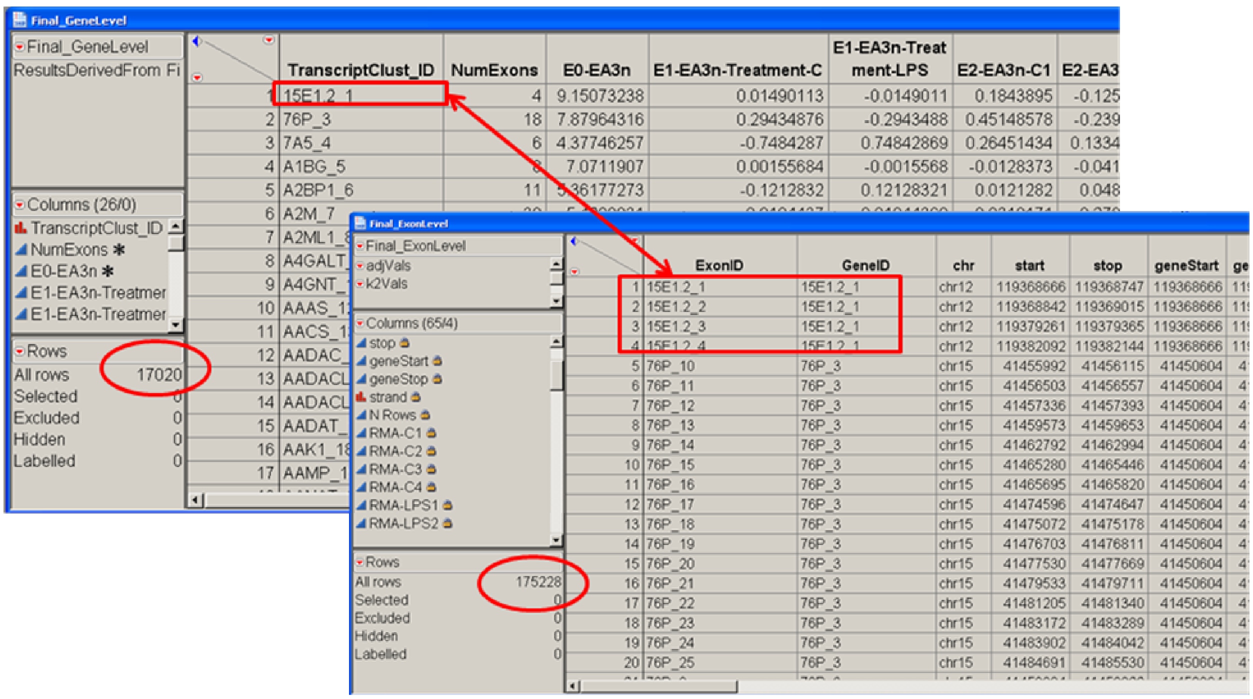
You will need to have created a Treatment column labeling the treatments in the study in the MasterFile.

You will also need to choose what type of Pre-Analysis filters under the "Exon Inclusion criteria" you would like to apply and whether the "Affymetrix Tissue dataset" will be used to look for Uniformly unresponsive probesets.





Final_GeneLevel and Final_ExonLevel files are printed out. These files are linked up by the "TransClustID" or "GeneID" columns.



ExonANOVAnested printed out Column prefix descriptions:

E0 - ANOVA mean
E1 - treatment effects(A term in ANOVA)
E2 - random effects for replication within tretament (Beta term in ANOVA)
E3 - exon (or probeset) effect (C term in ANOVA)
E13 - treatment-exon interaction effects (AC term in ANOVA)
yAdj - Expression data for each sample, adjusted for exon (or probeset) effect
yFit - Mean of each treatment group of yAdj
maxABSInteractionEffect - maximum absolute value E13, over all interaction effects for a given gene
p-EA3n-Treatment - p-value for treatment effect (the A term in ANOVA)
p-EA3n-sample(Treatment) - p-value for the replicate within treatment effects, E2
p-EA3n-exon - p-value for the exon (or probeset) effect (C term in ANOVA)
p-EA3n-TreatmentXexon - p-value for the treatment by exon interaction effect (AC in ANOVA)

Where “Treatment” is the name of the treatment column in the MasterFile